Roger Penrose
Collected Works

VOLUME 3

1976–1980

OXFORD SCIENCE PUBLICATIONS

OXFORD
UNIVERSITY PRESS

www.oup.com

ISBN 978 0 19 921936 4

ROGER PENROSE
COLLECTED WORKS

VOLUME 3

1976–1980

常州大学图书馆
藏书章

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP
Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto
With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam
Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries
Published in the United States
by Oxford University Press Inc., New York
© Roger Penrose 2011
The moral rights of the authors have been asserted
Crown copyright material is reproduced under Class Licence
Number C01P0000148 with the permission of OPSI
and the Queen's Printer for Scotland
Database right Oxford University Press (maker)
First published 2011
All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system, or transmitted, in any form or by any means,
without the prior permission in writing of Oxford University Press,
or as expressly permitted by law, or under terms agreed with the appropriate
reprographics rights organization. Enquiries concerning reproduction
outside the scope of the above should be sent to the Rights Department,
Oxford University Press, at the address above
You must not circulate this book in any other binding or cover
and you must impose the same condition on any acquirer
British Library Cataloguing in Publication Data
Data available
Library of Congress Control Number: 2010926298
Typeset by Newgen Imaging Systems Pvt Ltd., Chennai, India
Printed in Great Britain
on acid-free paper by
CPI Antony Rowe, Chippenham, Wiltshire
ISBN 978-0-19-921938-4 (Vol. 3)
978-0-19-921944-5 (Set)
1 3 5 7 9 10 8 6 4 2

PREFACE

The topics addressed, in these *Collected Works*, vary over quite an extensive range, the boundaries between their areas being sometimes quite blurred. It would have been hard to classify all these articles and arrange them according to individual topics. Moreover, my own mode of working has been, on the whole, to keep most of my different scientific or mathematical interests running concurrently. Accordingly, it has seemed appropriate not to attempt to group the papers together into separate tidy categories. To do so would have provided a somewhat misleading picture of my overall mode of thinking. Instead, the ordering of articles that has been adopted is essentially chronological, at least with regard to the date of publication. The resulting somewhat higgledy-piggledy appearance that we find in the sometimes abrupt juxtaposition of different topics is, perhaps, fairly illustrative of my actual mode of working.

There are two main kinds of drive that underlie these works. On the one hand, some of the articles have arisen mainly from purely mathematical considerations, which can be ultimately aesthetic or just matters of simple curiosity. On the other hand, motivations for many others of the articles came primarily from a strong desire to discover more about the underlying principles governing the workings of the physical world. Yet, one of the deep mysteries about what has been discovered about these principles is the great breadth of intrinsically appealing mathematics that is found to underlie the actual operation of the physics of the universe, so these two drives are found to relate closely to one another. The reader may perhaps note the distinct geometrical flavour of much of my own mathematics, evidencing the strong appeal that geometry exerts on me, an appeal that is not infrequently intertwined with desire for algebraic elegance. Indeed, I have found it remarkable how much of the workings of our universe appear to depend upon sophisticated and powerful geometry and algebra.

My own very geometrical perspectives, as opposed to purely computational ones, are perhaps made more obvious in the later volumes of this series than in the early ones, but this appearance is deceptive. Very early in my career as a Cambridge graduate student, I had developed a graphical notation which enabled me to carry out many kinds of calculation in an entirely diagrammatic way. For a long time I was rather secretive of this procedure, and in my earlier writings would take care to remove all trace of these visual origins. I do not doubt that my predilections for such geometric or algebraically elegant forms of understanding have influenced the specific choices of direction in which my endeavours towards furthering our understandings of the physics of the world have taken.

The focus of these directions has become more clarified as time has moved on, and has narrowed down to such a degree that there has emerged a somewhat disturbing repetitiveness in many of the later articles. A good number of these provide rather non-technical accounts. An example of such repetitiveness occurs with my descriptions of gravitational collapse to a black hole. There is, however, a gradually developing outlook on this issue, leading to an emphasis on the fundamental contrast in character between the space-time singularities that inevitably appear within black holes and the space-time singularity that constitutes the big bang, this contrast being central to the existence of a second law of thermodynamics in our universe. The same kind of repetitiveness applies to my frequently repeated claim of a need for a fundamental change in the basic laws of quantum theory, where only towards the end do my suggestions become particularly specific. Likewise, many mildly differing accounts are provided outlining the formalism and the basic ideas that underlie twistor theory. The later ones lead up to various viewpoints on the 'googly' problem, which has constituted the main obstruction to the finding of a full understanding of general relativity in twistor terms. Despite such repetition, it is my hope that there is a sufficient variety of emphasis and of historical development in such articles for it to have been worth while to include them all here.

Several colleagues have collaborated with me in writing the works presented here, and I can mention but only a few. Foremost, I have collaborated long and particularly fruitfully with Ezra T. Newman, especially with regard to the asymptotics of general relativity, and also with Wolfgang Rindler in the development of the formalism of 2-spinor calculus. Other collaborators who have played important roles in helping to develop the mathematics of discrete physics and twistor theory have been Erwin Kronheimer, Michael Eastwood, Ronny Wells, and Denny Hill. Among my collaborators, providing very significant inputs, who were former students in my research group, are Andrew Hodges, Paul Tod, Lane Hughston, George Sparling, Richard Ward, Nick Woodhouse, and Lionel Mason, but there were many others. Although not playing roles as actual collaborators, fundamental influences in my research have come from Dennis Sciama and Michael Atiyah.

Roger Penrose
April 2010

BIOGRAPHY

Early years and influences

I was born in Colchester, Essex, UK on 8 August 1931. My father, Lionel Sharples Penrose, FRS, was a highly talented and intelligent man, with many interests. He, himself, had come from a Quaker family, his own father (James Doyle Penrose) having been a professional artist and his mother (Josephine Peckover) was from a well-to-do Quaker banking family. Lionel's primary professional interest was in the causes of hereditary mental disease. He was Galton Professor of Human Genetics at University College London from 1945 and, following retirement, Director of the Kennedy–Galton Institue, St Albans. He had a very considerable artistic talent and a keen interest in music, especially Mozart, but also Bach and other classical composers, dabbling a little in composition himself, particularly when there was a puzzle aspect in what was involved (where he once wrote a piece that could be played equally well upside down as right-way up). He liked making puzzles of various kinds in wood, and spent much of his later life designing wooden models illustrating self-reproduction. He was recognized as a fine composer of chess problems. He had a definite fondness for mathematics and was an innovative statistician. He loved explaining matters of science and mathematics. On one occasion, when I told him that my mathematics teacher at school had informed us that he would be explaining calculus the next day, Lionel immediately took me aside so as to be sure to be the first to teach me about the beauties of calculus. This made a great impression on me. Yet he was an extremely difficult person to communicate with on personal matters.

My mother, born Margaret Leathes, was a charming and caring woman with a very subtle mind. She had been Head Girl at Bedales school, and became medically qualified, but found she was not able to practice after her marriage. After Lionel died, in 1972, she married Max Newman FRS, a fascinating man, who had been Professor of Mathematics in Manchester and, as we later learned, a key figure at Bletchley Park during the war. Margaret's father, John Beresford Leathes, FRS, had been a distinguished professor of physiology, and her mother, a concert pianist, born Sara Mara Natanson from a Jewish family in Latvia, had disowned her family (then in St Petersberg), becoming completely secretive about her origins, and even about her maiden name. Margaret enjoyed geometry at school, but did not

communicate much on this topic with me. My older brother Oliver was very precocious, and I learnt a lot of physics and mathematics from him as I was growing up. He later became a distinguished mathematical physicist (statistical mechanics) and FRS. My younger brother Jonathan excelled at games, particularly chess, and later became British Chess Champion a record 10 times, and then a top correspondence chess player. My much younger sister Shirley Victoria, now Shirley Hodgson, became a highly regarded Professor of Cancer Genetics.

This atmosphere of science, mathematics, art, and music, and of puzzles and games, had been an important part of my upbringing. During the war years, the family lived in London, Ontario, Canada, where Lionel became Director of Psychiatric Research at the Ontario Hospital. We returned to England in 1945, where I attended University College School, London. As an undergraduate, I studied Mathematics at University College London, from 1951 to 1954; then I went to St John's College Cambridge to do research in algebraic geometry, initially under William Hodge and subsequently John Todd. In my same year, under Hodge, was Micheal Atiyah, who was an inspirational, though somewhat intimidating colleague. Frank Adams and Christopher Zeemen (see Chapter 13) were contemporaries who also influenced me, and I became good friends with Hans Liebeck and Douglas Munn (see Chapter 3), who were close research-student colleagues at St John's.

Yet, the colleague who had the greatest influence on the development of my research during this period, and for many years later, was the cosmologist Dennis Sciama. Even before studying at Cambridge, I had a chance meeting with him during a visit to my brother Oliver in Cambridge in which I raised a question that had worried me in relation to the inspiring series of BBC radio talks given by Fred Hoyle in 1951 (The nature of the universe. A series of broadcast lectures). On the strength of this encounter (at which I used a space-time diagram with tilted light cones to make my point), Dennis had evidently formed the opinion that it would be worth while to interest me more in physics, even trying, later, to persuade me to switch my topic of research to cosmology. I was then too committed to my many pure-mathematical activities, but I learnt a great deal of physics from Dennis, and of the excitements involved in striving to uncover secrets of the universe. Another influential colleague was Felix Pirani, from whom I learnt much about developments in general relativity.

Pivotal were the superb lecture courses of Bondi, on general relativity, and Dirac, on quantum mechanics. On the pure-mathematical side, I recall particularly an elegant information-packed algebra course by Philip Hall. Lectures by Steen on mathematical logic turned out to influence me strongly towards my later viewpoint concerning the relation between computation and conscious understanding (e.g. Chapters 169, 184, 186).

Degrees

1952 BSc Mathematics (Spec) (1st Cl), University College, London

1957 PhD (Cantab), St John's College, Cambridge

Employment

1956	NRDC London (research on numerical analysis with computers); later consultant
1956–1957	Assistant Lecturer, Bedford College, London
1957–1960	Research Fellow, St John's College, Cambridge
1959–1961	NATO Research Fellow, Princeton, Syracuse, Cornell Universities
1961–1963	Research Associate, King's College, London
1963–1964	Visiting Associate Professor, University of Texas, Austin, Texas
1964–1966	Reader (Applied Mathematics), Birkbeck College, London
1966–1967	Visiting appointments: Yeshiva, Princeton, Cornell, Chicago Universities; lecturer Battelle Inst., Seattle
1967–1973	Professor (Applied Mathematics), Birkbeck College, London
1969–1970	Visiting professor, Yeshiva New York, Princeton, and Cornell Universities
1973–1998	Rouse Ball Professor (Mathematics), Oxford; 1998– Emeritus
1978	Miller Fellow, Univ. of California, Berkeley
1982–1987	Lovett Professor (Mathematics), Rice Univ., Houston, Texas, USA (Part-time)
1987–1993	Distinguished Professor (Physics, Mathematics), Syracuse Univ. USA (Part-time)
1993–	Francis and Helen Pentz Distinguished Professor (Physics, Mathematics; part-time)
1998–2001	Gresham Professor (Geometry), Gresham College, London

Honours Prizes include 1971 Heinemann Prize; 1972 Fellow of Royal Society; 1978 Plenary Lecture, International Congress of Mathematicians, Helsinki; 1985 Royal Medal (Royal Society);1988 Wolf Prize (physics, with Hawking); 1990 Albert Einstein Medal; 1994 knighted (services to science); 1994 Polish Academy of Sciences; 1998 US National Academy of Sciences; 2000 Order of Merit; 2004 de Morgan Medal (London Mathematical Society); 2001 Royal Irish Academy (Science); 2008 Copley Medal (Royal Society).

Marriages

1958	Joan Isabel Wedge; divorced 1980
1988	Vanessa Dee Thomas

Children

1963	Christopher Shaun
1964	Toby Nicholas
1966	Eric Alexander
2000	Maxwell Sebastian

Honorary degrees

1992	University of New Brunswick, Canada
1993	University of Surrey
1994	University of Bath
1995	University of London
1996	University of Glasgow
1996	University of Essex
1997	University of St Andrews
1998	University of Santiniketan, India
1998	Open University
2002	University of Southampton
2004	University of Waterloo, Canada
2005	University of Leuven, Belgium
2005	Warsaw University
2005	University of Athens
2006	University of York

Books

1972	Techniques of Differential Topology in Relativity (NSF and SIAM)
1984 and 1986	Spinors and Space-Time (Volumes 1 and 2; Cambridge University Press; with Rindler)
1989	The Emperor's New Mind (Oxford University Press 1989)
1994	Shadows of the Mind (Oxford University Press)
1996	The Nature of Space and Time (Princeton University Press; with Hawking)
1997	The Large, the Small and the Human Mind (Cambridge University Press)
2004	The Road to Reality (Jonathan Cape)

INTRODUCTION

Important new realizations concerning twistor theory came to our Oxford twistor group during the period covered by this volume. There were two main sources from earlier work, but then the ideas came together to give us a deeper understanding of what is involved, providing us with a new perspective on the way that mathematical features of the complex geometry of twistor theory relate to actual physical fields.

One of these sources came from the striving to find a good definition of total angular momentum for an asymptotically flat space-time. The problem was to find a satisfactory intrinsic way to define the notion of an *asymptotic rotation* which is 'supertranslation-free', following a burst of outgoing gravitational radiation (see Chapter 26). In the absence of such a definition, the normal procedures (e.g. using Noether's theorem) do not seem to work. In flat Minkowski space $\mathbb{M}$, a genuine rotation is recognized asymptotically by the fact that it leaves invariant a family of 'good' (i.e. supertranslation-free) cross-sections ('cuts') of null conformal infinity $\mathscr{I}^+$, these being intersections of ordinary light cones with $\mathscr{I}^+$. When gravitational radiation is present, the characterization of such a 'light cone' as a null hypersurface with a proper vertex does not work, but one can attempt a different characterization that works for $\mathbb{M}$, namely that the good cuts have the property that the asymptotic outgoing shear σ vanishes all along the cut. This would have been feasible were it not for the fact that σ is a *complex* quantity, and there is not, in general, the real-number freedom to arrange for $\sigma = 0$ over the entire S^2 that represents the cut. Ted Newman then proposed what turned out to be a remarkably insightful idea, namely to allow a *complex* supertranslation to be applied, so that the cut becomes a complex (holomorphic) cross-section of the complexified null infinity $\mathbb{C}\mathscr{I}^+$. The space of Newman's *good cuts*, using a curious (but, on reflection, natural) definition of complex metric that he introduced, was a complex-Riemannian 4-space that he referred to as $\mathcal{H}$-space ('$\mathcal{H}$' standing for 'Heaven'—i.e. 'where the good Cohens [cones] go'). This is described in Chapter 97. When I finally realized what Ted was actually doing, it became evident to me that his 'good cuts', in this sense, must be represented as holomorphic curves (Riemann spheres) in asymptotic twistor space (see Chapter 52). It also turned out that $\mathcal{H}$-space is automatically 'empty'—i.e. *Ricci-flat*[1] and, with my own definitions of chirality, etc., *anti-self-dual*. In twistor language, this was essentially the 'non-linear graviton construction' asserted in Chapters 67 and 71 and made explicit in Chapter 70.

[1] I was away at a conference in Italy, I believe, when I had persuaded myself that I could demonstrate this fact, and I puzzled the people in the conference office by getting them to send Ted a telegram with the cryptic message 'HEAVEN IS EMPTY'.

The other source of input came from the pure mathematicians whom I consulted about the issue of when one could be assured that the appropriate holomorphic curves actually exist as required for the construction. I initially consulted Isadore Singer about this, while I was away in the USA and he gave me an appropriate criterion. However, the language in which he phrased it involved the partial differential equations of a Dolbeault formulation, about which I could form little intuition, so I tried again with Michael Atiyah after my return to Oxford, and his approach—more along Čech lines, and involving the projective geometry that I was familiar with—was more to my liking and I was then able to comprehend the Kodaira theorems that were needed for the non-linear graviton construction (Chapters 70 and 97). This led to a further educational sessions with Atiyah involving my whole research group, and we began to realize how many of the puzzling aspects of twistor theory could be understood in terms of complex sheaf cohomology and complex-manifold theory. Michael Eastwood joined us and our association with him proved invaluable.

Very soon after I had found this non-linear graviton construction, Richard Ward, who was a graduate student with us at the time, quickly saw how to adapt these kinds of idea so as to find a corresponding type of twistor construction for (anti-)self-dual electromagnetism, which he rapidly generalized so that the general (anti)-self dual Yang–Mills field could be correspondingly obtained. Michael Atiyah again got involved and the subject very much took off. In relation to these developments for non-linear fields, we realized the importance of educating ourselves in many of the underlying concepts needed for the proper understanding of complex manifolds, and we were enormously fortunate that Michael Atiyah was at hand in Oxford to explain these things to us. Most important were basic notions of holomorphic sheaf cohomology, and it gradually became clear, as a result of these conversations that it was in terms of such notions that many of the puzzling features of twistor contour integrals were to be understood. The theory of integrable systems gained much from these developments, most specifically from Richard Ward's construction (see Ward, R.S. (1977) On self-dual gauge fields, *Phys. Lett.* **61A**, 81–2; Mason, L.J. and Woodhouse, N.M.J. (1996) *Integrability, Self-Duality, and Twistor Theory*, Oxford University Press, Oxford).

During this period of fruitful activity in twistor theory, made largely by members of our Oxford research group, but a number of others were involved also, we founded the informal publication *Twistor Newsletter* (often abbreviated $\mathbb{TN}$), produced at the Mathemtical Institute, Oxford. Roughly two volumes appeared per year, with articles, often handwritten, of varying degrees of technicality, seriousness, readability, reliability, and depth. These continued until shortly after my official retirement in 1998. A good proportion of the articles have been reprinted in four publications, with titles [*Further*] *Advances in Twistor theory*, published by Pitman, Pergamon, and Chapman and Hall. Often, members of our group, or other contributors, would be referred to in the articles merely by their initials, so to help in identifying these people, here is a partial list, with the last name spelled out: AHelfer, APHodges, APilato, BDBramson, DHill, EJDunne, GAJSparling, GBurnett-Stuart, KPTod, LJMason, LPHughston, MASinger, MFAtiyah, MGEastwood, NJeffaryes, NJHitchin, NMJWoodhouse, RJBaston, RMBryant, ROWells, RPenrose, TNBailey, TsouS-T, VDThomas, WSShaw, ZPerjés.

A paper of mine was finally published after some 15 years, as a 'golden oldie' (Chapter 114). This was concerned with the characteristic initial value problem, both in flat space-time (where I had employed a generalization of the Kirchoff–D'Adhemar integral formula that I had previously found) and for general relativity. The idea of exact sets described in Chapter 12 was central to the arguments

Finally, a mention should be made of the 2-piece aperiodic tile shape sets described in Chapter 111.

ACKNOWLEDGEMENTS

Oxford University Press has attempted to trace the copyright holder of all the papers included in these volumes and apologizes to copyright holders if permission to publish in this form has not been obtained.

The author and publishers thank the following individuals for their permissions:

Acta Physics Polonica B **280**
Adam Hilger Ltd **147**
American Chemical Society **206**
American Institute of Physics
American Journal of Physics **19, 24**
Journal of Maths and Physics **15, 26, 27, 29, 39, 41, 58, 84**
Errata **15**
American Mathematical Society **30, 180**
Bulletin of the American Mathematical Society **144**
Contemporary Mathematics **179**
American Physical Society
Physical Review Letters **16, 20, 23, 302**
Reviews of Modern Physics **21**
Annals of Human Genetics **100**
Asian Journal of Mathematics **281**
Astronomische Gesellschaft **295**

Benjamin **36**
Bibliopolis **175**
Blackwell **80, 166, 186, 215**
Annals of the New York Academy of Sciences **60, 190, 294**
British Broadcasting Company **59**
British Journal of Psychology **9**

Cambridge Entomological Club **243**

Cambridge University Press **6, 7, 46, 104, 138, 157, 172, 173, 176, 195, 218, 220, 227, 230, 240, 290, 303, 304**

Behavioural and Brain Sciences **193, 194, 224**

Proceedings of the Cambridge Philosophical Society **3, 4, 5, 32, 42, 50, 126**

Chapman and Hall/CRC Press **107, 124, 125, 127, 128, 130, 136, 139, 142, 143, 146, 149, 150, 154, 155, 156, 160, 161, 164, 170, 171, 178, 189, 191, 198, 201, 203, 209, 229, 234, 293**

Connected Globe **257**

Consigle Nazionale Della Richerche **145**

Cornell University Press **33**

Éditions du Centre National de la Recherche Scientifique **14**

Elsevier **12, 45, 52, 68, 85, 86, 92, 93, 123, 151, 167, 185, 187, 226, 279**

Escher, M.C. images: The M.C. Escher Company B.V. **218, 249, 306**

Eureka **1, 2**

Folio Society **284**

Freeman **56**

Granta **297**

Gresham College **277**

Halsted Press **148**

Hanser **66, 67**

Imperial College Press **264**

Imprint Academic **235, 245**

Indian Academy of Sciences **244**

Indiana University Press **25**

Institute of Electrical and Electronics Engineers **196**

Institute of Physics **217**

Classical and Quantum Gravity **253, 254, 274, 282**

European Journal of Physics **258**

Journal of Physics A: Mathematical and General **95**

International Astronomical Union

International Astronomical Union Symposium Proceedings Series No. 13, Cosmology **72**

International Atomic Energy Agency
Contemporary Physics **38**
International Centre for Theoretical Physics **53, 68**
International Mathematical Union **286**
International University Bremen **305**
Inter-University Centre for Astronomy and Physics **265**
The Italian Astronomical Society
Memorie della Società Astronomica Italiana **200**

Longman **78, 79, 108, 116, 117, 134, 135, 141, 177, 188**

The Massachusetts Institute of Technology Press **228, 246**

National Institute of Sciences of Ukraine **48**
Nature **8, 44, 49**
Nature (Physical Science) **47**
New Scientist **10, 101**
New York Academy of Sciences **294**
Oxford University Press **55, 69, 121, 122, 165, 178, 184, 197, 202, 208, 216, 237, 270, 272, 276, 291, 299**

Pearson **109**
Penguin **273**
Penrose's Unpublished Papers **31**
Pergamon Press **17**
Pitman **73, 75, 76, 77, 82, 87, 89, 90, 94, 102, 103, 110**
Pontificial University **296**
Princeton University Press **28, 268**
Annals of Math **13, 131**
Proceedings of the International Congress of Mathematicians **115**

Rivista del Nuovo Cimento **40**
Routledge **169**
Royal Astronomical Society **183**
The Royal Society
Proceedings of the Royal Society Series A **22, 35, 97, 105, 106, 132**
Philosophical Transactions of the Royal Society of London. Series A **275, 283**

The Sciences **99**
Scientific American **51**
Simon and Schuster **242**
Springer **61, 63, 65, 83, 91, 98, 111, 133, 168, 204, 207, 231, 255, 263, 278, 292, 298, 300**
Artificial Intelligence Review **205**
Communications in Mathematical Physics **43, 54, 129**
Cosmology and Gravitation **120**
Foundations of Physics **140**
General Relativity and Gravitation **70, 71, 113, 114, 239, 251**
Higher-Order and Symbolic Computation **287**
International Journal of Theoretical Physics **37, 56**
Journal of Statistical Physics **233**
Lecture Notes in Physics **112, 152**
The Mathematical Intelligencer **285**
Taylor and Francis **18**
International Studies in the Philosophy of Science **256**
Surveys in High Energy Physics **119**
Trinity College, Dublin **219**

World Scientific **162, 163, 225, 236, 241, 252, 289, 301**

Universal Academy Press **266**
University of Chicago Press **96, 269**
University of Chittagong **158**
University College London **267**

The University of Oxford (Mathematics Department)
Twistor Newsletter **73, 74, 75, 76, 77, 78, 79, 81, 82, 88, 89, 90, 94, 102, 103, 107, 108, 116, 117, 118, 124, 125, 127, 128, 130, 134, 135, 136, 137, 139, 141, 142, 143, 146, 149, 150, 153, 154, 155, 156, 157, 160, 161, 164, 170, 171, 174, 181, 182, 188, 189, 191, 192, 198, 199, 201, 203, 209, 210, 211, 212, 213, 214, 221, 222, 223, 229, 232, 238, 248, 290, 250, 259, 260, 261, 262, 271, 288**

University of Utah Press **247**

CONTENTS
VOLUME 3

Introduction *xxiii*

1970s

70 Nonlinear gravitons and curved twistor theory 1
71 The nonlinear graviton 25
72 The space-time singularities of cosmology and in black holes 33
73 The twistor quadrille 45
74 The snail contour 53
75 The universal bracket factor 55
76 Some elementary twistor integrals with boundary 59
77 Twistor functions and sheaf cohomology 63
78 How to grow (or at least half grow) new twistors from old 77
79 Some remarks on the two twistor description of electromagnetic and linearized gravitational potentials 81
80 The cosmic censorship hypothesis 83
81 A finite covering for PT^+ 87
82 The back-handed photon (or, to a cricketer, the googly photon!) 89
83 Any space-time has a plane wave as a limit 99
84 The Kähler structure of asymptotic twistor space 105
85 The twistor programme 113
86 Space-time singularities 127
87 The twisted camel (or how to get through the eye of a singularity). Includes: more on the twisted camel 137
88 Remarks concerning the ambidextrous photon 145
89 Massless fields and sheaf cohomology 149
90 Note on the n-twistor internal symmetry group 159
91 Aspects of general relativity 163
92 Is nature complex? 203
93 Gravitational collapse: A review 213
94 Local H^1's and propagation 231
95 Extended Regge trajectories 237

96	Singularities of spacetime	243
97	The metric and curvature properties of $\mathcal{H}$-space	271
98	The geometry of the universe	297
99	Einstein's vision and the mathematics of the natural world	341
100	The topology of ridge systems	347
101	Twisting round space-time	359
102	A googly graviton?	365
103	Twistors as helicity raising operators	375
104	Singularities and time-asymmetry	381
105	General introduction	441
106	Global general relativity	445
107	Remarks on the Sparling and Eguchi-Hanson (googly?) gravitons	449
108	The anti-self-dual Coulomb field's non-Hausdorff twistor space	459
109	On the twistor descriptions of massless fields	461
110	Combinatorial quantum theory and quantized directions	499
111	Pentaplexity: A class of nonperiodic tilings of the plane	507

1980s

112	Recent advances in global general relativity: A brief survey	515
113	Twistors for flat and curved space-time	527
114	Null hypersurface initial data for classical fields of arbitrary spin and for general relativity	575
115	The complex geometry of the natural world	617
116	On the evaluation of twistor cohomology classes	625
117	Three channels for the box diagram	629
118	Puzzle: A sequence for twistorians	637
119	A brief introduction to twistors	639
120	A brief outline of twistor theory	663

DETAILED CONTENTS
VOLUMES 1–6

VOLUME 1

1950s

1 Morley's trisector theorem 1
2 'Construction problem' and 'Word problem' 3
3 A note on inverse semigroups 5
4 A generalized inverse for matrices 11
5 On best approximate solutions of linear matrix equations 21
6 Tensor methods in algebraic geometry 25
7 Geometrical algebras: A new approach to invariant theory 251
8 A self-reproducing analogue 351
9 Impossible objects: A special type of visual illusion 355
10 Puzzles for Christmas 359
11 The apparent shape of a relativistically moving sphere 365

1960s

12 A spinor approach to general relativity 369
13 Imbedding of manifolds in Euclidean space 403
14 General relativity in spinor form 415
15 An approach to gravitational radiation by a method of spin coefficients 421
16 Asymptotic properties of fields and space-times 437
17 The light cone at infinity 441
18 Conformal treatment of infinity 447
19 Energy conservation as the basis of relativistic mechanics 469
20 Gravitational collapse and space-time singularities 475
21 A remarkable property of plane waves in general relativity 479
22 Zero rest-mass fields including gravitation: Asymptotic behaviour 487
23 10 exact gravitationally conserved quantities 533
24 Energy conservation as the basis of relativistic mechanics II 537
25 General-relativistic energy flux and elementary optics 541
26 Note on the Bondi-Metzner-Sachs group 559
27 Some new gravitationally conserved quantities 569
28 An analysis of the structure of space-time 579
29 Twistor algebra 731
30 Conserved quantities and conformal structure in general relativity 755
31 Theory of quantized directions 769
32 On the structure of causal spaces 801

VOLUME 2

1960s

33 Cosmological boundary conditions for zero rest-mass fields 1
34 On gravitational collapse and cosmology 15
35 New conservation laws for zero rest-mass fields in asymptotically flat space-time 29
36 Structure of space-time 61
37 Twistor quantization and curved space-time 177
38 On gravitational collapse 217

39 Solutions of the zero-rest-mass equations 231
40 Gravitational collapse: The role of general relativity 235
41 Conserved quantities in the Einstein–Maxwell theory 261

1970s

42 The singularities of gravitational collapse and cosmology 267
43 On quadratic first integrals of the geodesic equations for type {22} spacetimes 289
44 Scattering of two impulsive gravitational plane waves 301
45 Applications of negative dimensional tensors 307
46 Angular momentum: An approach to combinatorial space-time 333
47 Extraction of rotational energy from a black hole 367
48 Spinor classification of energy tensors 375
49 Black holes and gravitational theory 389
50 Ideal points in space-time 395
51 Black holes 419
52 Twistor theory: An approach to the quantisation of fields and space-time 429
53 a) 'Radiation, energy and the BMS group'; b) 'Some recent developments'; c) 'N-P constants' 505
54 On a quadratic first integral for the charged particle orbits in the charged Kerr solution 513
55 The geometry of impulsive gravitational waves 521
56 On the nature of quantum geometry 537
57 Internal instability in a Reissner–Nordström black hole 583
58 A space-time calculus based on pairs of null directions 599
59 Black holes 609
60 Naked singularities 635
61 Relativistic symmetry groups 647
62 Gravitational collapse 707
63 Gravitational collapse 729
64 The role of aesthetics in pure and applied mathematical research 741
65 Singularities in cosmology 749
66 Twistors and particles: An outline 761
67 The non-linear graviton 779
68 Classical gravity: A status report 791
69 Twistor theory: Its aims and achievements 795

VOLUME 3

1970s

70 Nonlinear gravitons and curved twistor theory 1
71 The nonlinear graviton 25
72 The space-time singularities of cosmology and in black holes 33
73 The twistor quadrille 45
74 The snail contour 53
75 The universal bracket factor 55
76 Some elementary twistor integrals with boundary 59
77 Twistor functions and sheaf cohomology 63
78 How to grow (or at least half grow) new twistors from old 77
79 Some remarks on the two twistor description of electromagnetic and linearized gravitational potentials 81
80 The cosmic censorship hypothesis 83
81 A finite covering for PT^+ 87
82 The back-handed photon (or, to a cricketer, the googly photon!) 89
83 Any space-time has a plane wave as a limit 99
84 The Kähler structure of asymptotic twistor space 105
85 The twistor programme 113
86 Space-time singularities 127
87 The twisted camel (or how to get through the eye of a singularity). Includes: more on the twisted camel 137
88 Remarks concerning the ambidextrous photon 145
89 Massless fields and sheaf cohomology 149
90 Note on the n-twistor internal symmetry group 159
91 Aspects of general relativity 163
92 Is nature complex? 203
93 Gravitational collapse: A review 213
94 Local H^1's and propagation 231

95 Extended Regge trajectories 237
96 Singularities of spacetime 243
97 The metric and curvature properties of $\mathcal{H}$-space 271
98 The geometry of the universe 297
99 Einstein's vision and the mathematics of the natural world 341
100 The topology of ridge systems 347
101 Twisting round space-time 359
102 A googly graviton? 365
103 Twistors as helicity raising operators 375
104 Singularities and time-asymmetry 381
105 General introduction 441
106 Global general relativity 445
107 Remarks on the Sparling and Eguchi-Hanson (googly?) gravitons 449
108 The anti-self-dual Coulomb field's non-Hausdorff twistor space 459
109 On the twistor descriptions of massless fields 461
110 Combinatorial quantum theory and quantized directions 499
111 Pentaplexity: A class of nonperiodic tilings of the plane 507

1980s

112 Recent advances in global general relativity: A brief survey 515
113 Twistors for flat and curved space-time 527
114 Null hypersurface initial data for classical fields of arbitrary spin and for general relativity 575
115 The complex geometry of the natural world 617
116 On the evaluation of twistor cohomology classes 625
117 Three channels for the box diagram 629
118 Puzzle: A sequence for twistorians 637
119 A brief introduction to twistors 639
120 A brief outline of twistor theory 663

VOLUME 4

1980s

121 Time-asymmetry and quantum gravity 1
122 Some remarks on twistors and curved-space quantization 31
123 On Schwarzschild causality—a problem for "Lorentz covariant" general relativity 47
124 Concerning a Fourier contour integral 61
125 A new angle on the googly graviton 65
126 A quadratic mapping with invariant cubic curve 73
127 Twistors for cosmological models 91
128 Physical left-right symmetry and googlies 97
129 Cohomology and massless fields 105
130 A theory of 2-surface ('superficial') twistors 153
131 Some unsolved problems in classical general relativity 161
132 Quasi-local mass and angular momentum in general relativity 201
133 Time-asymmetry, cosmological uniformity and space-time singularities 213
134 Real-space topology for advanced/retarded twistor functions 217
135 Twistor blisters and relative cohomology 221
136 On the geometry of googly maps 225
137 A puzzle for twistorians 229
138 Some remarks on gravity and quantum mechanics 231
139 General-relativistic kinematics? 241
140 Spinors and torsion in general relativity 247
141 Non-locality in physics: A new mathematical viewpoint 263
142 On closed-set coverings for PT^+—A correction 269
143 Remarks on curved-space, twistor theory and googlies 271
144 A note on Sparling's 3-form 277
145 Physical space-time and nonrealizable CR-structures 281
146 Quasi-local mass and angular momentum 305
147 Relative cohomology, googlies and deformations of $\mathbb{I}$ 309
148 Donaldson's moduli space: A 'model' for quantum gravity? 313
149 Mass in general relativity 319
150 'New improved' quasi-local mass, and the Schwarzschild solution 331
151 On Bryant's condition for holomorphic curves in CR space 339

152 Integrals for general-relativistic sources: A development from Maxwell's electromagnetic theory 343
153 Mass and angular momentum at the quasi-local level in general relativity 377
154 Objective state-vector reduction? 387
155 Embedding 2-surfaces in CM 393
156 A suggested further modification to the quasi-local formula 397
157 Higher-dimensional two-surface twistors 399
158 Numerical relativity by power series 403
159 Elements of twistor theory 409
160 Energy and its definition in general relativity 429
161 Two-surface twistors and hypersurface twistors 441
162 Exponentiating a relative H^2 445
163 Gravitational field mass 449
164 Gravity and quantum theory 467
165 Twistors and the time-irreversability of state-vector reduction 473
166 Gravity and state vector reduction 477
167 Big bangs, black holes and 'time's arrow' 497
168 Escher and the visual representation of mathematical ideas 525
169 Hermann Weyl, space-time and conformal geometry 541
170 Quantum physics and conscious thought 573
171 Local twistor transport at $\mathscr{I}^+$: An approach to the googly 601
172 An approach to a coordinate-free calculus at $\mathbb{I}$ 609
173 Twistors in general relativity 613
174 Newton, quantum theory and reality 633
175 On Qadir's intertwining operators 667
176 On the origins of twistor theory 671
177 New directions in quantum gravity 693
178 Pretzel twistor spaces 703
179 Twistors and state-vector reduction 713
180 Aspects of quasi-local angular momentum 717
181 Fundamental asymmetry in physical laws 727
182 Topological QFT and twistors: Holomorphic linking 741
183 Holomorphic linking: Postscript 745
184 Singularities and big-bang cosmology 747
185 On the physics and mathematics of thought 751
186 A twistor conformal field theory for four space-time dimensions 785
187 Minds, machines and mathematics 791
188 Tilings and quasi-crystals: A non-local growth problem? 811
189 Non-Hausdorff Riemann surfaces and complex dynamical systems 839
190 Using spinors in general relativity 845
191 A twistorial approach to the full vacuum equations 851

VOLUME 5

1990s

192 Difficulties with inflationary cosmology 1
193 Light rays near i^0: A new mass-positivity theorem 19
194 Extended Regge trajectories: 12 years after 27
195 Précis of *The Emperor's New Mind: Concerning computers, minds and the laws of physics* 29
196 Author's response – the nonalgorithmic mind 43
197 Twistor theory after 25 years – its physical status and prospects 59
198 Twistors, particles, strings and links 89
199 Mass positivity from focussing and the structure of space-like infinity 101
200 Twistor theory for vacuum space-time: A new approach 107
201 What does the big bang tell us about quantum gravity? 113
202 Twistors as spin 3/2 charges 123
203 Twistors as charges for spin 3/2 in vacuum 133
204 The mass of the classical vacuum 143
205 Twistors as spin 3/2 charges continued: SL(3,C) bundles 151
206 Response to Tony Dodd's 'Godel, Penrose and the possibility of AI' 159
207 The biggest enigma 161
208 $\mathcal{H}$-space and twistors 167
209 'Penrose and mathematical ability' 189

210	The most general {2,2} self-dual vacuum: A googly approach	193
211	On impulsive gravitational waves	199
212	Kinking and causality	205
213	Puzzle page	213
214	Solution to puzzle in TN34	215
215	Gravitational collapse of the wave function: New thoughts	221
216	Setting the scene: The claim and the issues	225
217	The modern physicist's view of nature	259
218	Gravity and quantum mechanics	311
219	Foreword to *Mr Tompkins in paperback*	323
220	Mathematics, physics, computers and conscious understanding	327
221	An example of indeterminacy in the time-development of "already unified field theory": A collision between electromagnetic plane waves	339
222	Orthogonality of general spin states	349
223	A simple observation concerning {2 2} vacuums	355
224	The orthographic image of a regular tetrahedron	357
225	Author's response – An Emperor still without mind	361
226	Square-root voting	369
227	On Bell non-locality without probabilities: More curious geometry	375
228	Quantum non-locality and complex reality	401
229	On the cohomology of impossible figures	415
230	Spin 3/2 fields and local twistors	419
231	Mathematical intelligence	427
232	Is conscious awareness consistent with space-time descriptions?	459
233	A twistor-topological approach to the Einstein equations	475
234	On the second law of thermodynamics	485
235	Twistors and the Einstein equations	491
236	What 'gaps'? Reply to Grush and Churchland	505
237	Non-locality and objectivity in quantum state reduction	521
238	Must mathematical physics be reductionist?	531
239	Concerning space-time points for spin 3/2 twistor spaces	547
240	In {2 2} vacuums, the null-datum on null rays is exactly r^{-5}	553
241	Why new physics is needed to understand the mind	557
242	The role of gravity in quantum state reduction	573
243	Consciousness involves noncomputable ingredients	593
244	Beyond the doubting of a shadow: A reply to commentaries on *Shadows of the Mind*	603
245	Chandrasekhar, black holes, and singularities	641
246	Conscious events as orchestrated space-time selections	661
247	Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness?	681
248	Space-time and cosmology	711
249	New ideas for the googly graviton	757
250	Incidence between complex null rays	771
251	Two puzzles	777
252	On gravity's role in quantum state reduction	779
253	Twistor theory, the Einstein equations, and quantum mechanics	801

VOLUME 6

1990s

254	Andrzej Mariusz Trautman	1
255	Twistor geometry of light rays	11
256	Remarks on tiling: Details of a $(1+\varepsilon+\varepsilon^2)$-aperiodic set	37
257	On understanding understanding	69
258	Mathematics: Pattern, precision and profundity	85
259	The mathematics of the electron's spin	109
260	Five things you can do with a surface in PT	115
261	Solutions to puzzles in TN41	121
262	Corrections to missprints in solutions to TN41 puzzles	125
263	Googly maps as forms	127
264	The need for a non-computational extension of quantum action in the brain	133
265	Chandrasekhar and the end of time	151
266	Science and the mind	161

267 Quantum state-reduction as a gravitational phenomenon; theory and experiment 165
268 Lionel Penrose: Colleague and father 169
269 Foreword to *Einstein's miraculous year: Five papers that changed the face of physics* 175
270 The question of cosmic censorship 185
271 Afterword to *The geometric universe: Science, geometry, and the work of Roger Penrose* 203
272 The twistor space for a strongly asymptotically flat vacuum 213
273 The complexity of our singular universe 223
274 Can a computer understand? 251
275 Spherically-symmetric solutions of the Schrödinger–Newton equations 279
276 Quantum computation, entanglement and state reduction 291
277 Foreword to *Ludwig Boltzmann: The man who trusted atoms* 305
278 Mathematics in the arts 309
279 Some remarks on twistor theory 323
280 The central programme of twistor theory 339
281 Lecture in honour of Leopold Infeld (extended outline only): Spinors in general relativity 371
282 On Ricci-flat twistor theory 381
283 Twistor theory and the Einstein vacuum 389
284 Introduction: Science at the turn of the millennium. 'Science into the next millennium: Young scientists give their visions of the future, part I.' 409

2000s

285 Introduction. *What is life?* 417
286 Drawing with complex numbers 433
287 Mathematical physics of the 20th and 21st centuries 441
288 Reminiscences of Christopher Strachey 459
289 On extracting the googly information 463
290 Wavefunction collapse as a real gravitational effect 489
291 On Bell non-locality without probabilities: Some curious geometry 509
292 The heritage of Pythagoras: Nineteen to the dozen 537
293 Twistors and general relativity 551
294 Towards a twistor description of general space-times: Introductory comments 579
295 Consciousness, the brain, and spacetime geometry: An addendum 597
296 The Schwarzschild singularity 605
297 The physics of the small and large: What is the bridge between them? 621
298 The rediscovery of gravity. The Einstein equation of general relativity 629
299 John Bell, state reduction, and quanglement 663
300 Consciousness, computation, and the Chinese room 677
301 Twistor geometry of conformal infinity 703
302 Gravitational collapse of the wavefunction: An experimentally testable proposal 713
303 Towards quantum superpositions of a mirror 719
304 On the instability of extra space dimensions 725
305 The problem of spacetime singularities: Implications for quantum gravity? 743
306 The remarkable role of mathematics in the physical universe 767

70
NONLINEAR GRAVITONS AND CURVED TWISTOR THEORY

The riddle of gravitation — on the occasion of the 60th birthday of Peter G. Bergmann (Proc. Conf., Syracuse Univ., Syracuse, N.Y., 1975). *General Relativity and Gravitation* **7** (1976), 1, 31–52

This article provides most of the detailed arguments underlying the 'non-linear graviton construction' of Chapter 67 (and 71)—although I did not spell out my somewhat awkward reasoning for the vanishing of torsion. The existence of this construction was, to me, a landmark, because it changed my view of how to incorporate general relativity into twistor theory. Previously, it has seemed to me to be likely that some sort of 'piecemeal' approach might be needed, perhaps something along the lines of Tullio Regge's simplicial approximations ('Regge calculus'), where flat space-time pieces are 'glued' together, and a standard twistor space would be assigned to each, and somehow quantum mechanics 'smoothed' the joins. But now we appear to have the potential to provide continuous (even analytic) curvature through gluing pieces of flat twistor space together in a non-linear way. See also Chapter 18. However, until a fully appropriate solution to the 'googly problem' is at hand (see Chapters 82, 102, 125, 136, 283, 289, 294), this procedure remains incomplete.

2 Nonlinear gravitons and curved twistor theory; The riddle of gravitation – on the occasion of the 60th birthday of Peter G. Bergmann (Proc. Conf., Syracuse Univ., Syracuse, N.Y., 1975). *General Relativity and Gravitation* **7** (1976), 1, 31–52

Nonlinear Gravitons and Curved Twistor Theory

ROGER PENROSE

Mathematical Institute, Oxford, England

§(1): *Introduction*

The question of how best to quantize gravity has been the subject of many discussions and arguments over the years. And Peter Bergmann has repeatedly and tirelessly reminded us that gravitational quanta should *not* be described in terms merely of *linearized* gravitation theory. I feel I have been rather slow at coming around to accepting this fully myself. It is, indeed, seductive to attempt to invoke the quantum-mechanical principle of linear superposition as an excuse for putting off, to a second stage of consideration, the complicated nonlinear nature of the gravitational self-interaction—and for putting off, perhaps indefinitely, the daunting encounter between quantum mechanics and the principles of curved-space geometry! If Peter Bergmann has taught us one thing above most others, it is surely that if we remove the life from Einstein's beautiful theory by steam-rollering it first to flatness and linearity, then we shall learn nothing from attempting to wave the magic wand of quantum theory over the resulting corpse.

Let me put things somewhat differently. Consider the common attitude according to which "gravitons" are described by linearized Einstein theory (spin-2 massless Poincaré covariant fields), a perturbative viewpoint being adopted starting from flat Minkowski space. If one such "graviton" is added to the vacuum (Minkowski) state the space remains flat. The null cones do not shift. If a second such "graviton" is added, and a third and a fourth, the space still remains flat, with null cones still locked in their original Minkowskian positions. With such a perturbative viewpoint it is only after an infinite number of "gravitons" have been added that the space can become curved. The situation may be compared with a power-series expansion. For example, with any finite

© 1976 Plenum Publishing Corporation, 227 West 17th Street, New York, N.Y. 10011. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, microfilming, recording or otherwise, without the written permission of the publisher.

number of terms the function

$$\frac{1}{z} + \frac{1}{z^2} + \cdots + \frac{1}{z^n}$$

has a pole singularity stuck at $z = 0$. But the sum to infinity

$$\frac{1}{z} + \frac{1}{z^2} + \frac{1}{z^3} + \cdots = \frac{1}{z-1}$$

has its pole shifted to $z = 1$. The usual perturbative expansions of graviton propagators have a similar behavior, with poles stuck to the Minkowskian null cones right up until the sum to infinity has been performed. Only then do the propagator singularities jump to their correct curved-space locations. But if it has meaning to talk about *physical* gravitons (see footnote 1), it should have meaning to talk about finite numbers of such gravitons. And if Einstein's theory is physically appropriate, then surely *each graviton itself must carry its measure of curvature*. I cannot believe that curvature emerges only after an infinite sum has been performed.

In the approach that I am suggesting here, a state consisting of just one graviton can be described, but in a way which involves both the curvature and nonlinearities of Einstein's theory. Though I feel that this approach can be justified solely on its own merits, it also receives striking encouragement from another direction: a surprising mathematical result enables one to construct the general such nonlinear graviton state from a *curved twistor space*, the construction being given in terms of one arbitrary holomorphic function of three complex variables. In this way, the approach fits naturally into the general twistor program [1–3] for the description of quantized fields.

§(2): *The Nonlinear Graviton*

In order to gain insights into how a graviton ought to be described, let us proceed, first, by analogy with the electromagnetic field. How, indeed, do we provide a mathematical description for a single free photon? There are, of course, many equivalent ways of doing this.

I shall here require a description which is both relativistically invariant and gauge invariant. The wave function can then be given as a solution $F_{ab}(= -F_{ba})$

[1] One might, on the other hand, adopt the view that a perturbational description in terms of (linear) "gravitons" is merely a convenience, no physical significance being attached to each individual "graviton." This might be legitimate if the program could be carried through successfully to the end. But although impressive progress has been made in recent years towards this goal [1], there remain fundamental problems whose solutions do not appear to be yet in sight. In particular, such standard perturbation techniques seem to lead, when sources are present, to a quantized Einstein theory which is not renormalizable [1].

Nonlinear gravitons and curved twistor theory; The riddle of gravitation – on the occasion of the 60th birthday of Peter G. Bergmann (Proc. Conf., Syracuse Univ., Syracuse, N.Y., 1975). *General Relativity and Gravitation* **7** (1976), 1, 31–52

of the free Maxwell equations

$$\nabla_{[a}F_{bc]} = 0 \qquad \nabla^a F_{ab} = 0 \tag{2.1}$$

However, there is an essential difference between a *classical* solution of the equations (2.1) and a solution giving the *wave function* of a *free photon*, for classical solutions are *real*, while wave functions are *complex*. Furthermore, a free photon wave function should be composed only of *positive frequencies*. There is also the question of normalizing the wave function, but that is not strictly necessary at this stage.

The requirement that a field in (compactified) Minkowski space M be of positive frequency can be stated neatly [4] in terms of the complexified (compactified) Minkowski space CM. Denote by CM^+ that region of CM whose points have complex position vectors

$$w^a = x^a - iy^a$$

(x^a and y^a being real), where y^a is *timelike* and points into the *future*; and denote by $\overline{\mathrm{CM}^+}$ the closure in CM of this set. Clearly we have $\mathrm{M} \subset \overline{\mathrm{CM}^+}$. A convenient condition to impose on a wave function defined (see footnote 2) on M, which ensures that it not only be of positive frequency but also normalizable, is that its domain of definition be extendable to the whole of $\overline{\mathrm{CM}^+}$ and that it be holomorphic (i.e. complex analytic) on this domain (where holomorphic on the boundary means extendable holomorphically just beyond the boundary). If we do not require normalizability and are prepared to allow the wave function to be discontinuous or a distribution (etc.) then we need only demand that its values on M be boundary values of some function holomorphic on CM^+.

Such a definition is much more convenient for our purposes than one which relies on a Fourier decomposition of the field. For general relativity does not lend itself easily to Fourier analysis. [Roughly, we can see the relation between the two approaches in the case of M as follows: the positive frequency Fourier component $\exp(-ip_a w^a)$ is damped down strongly in the region CM^+ when p_a is future-pointing and builds up in that region when p_a is past-pointing. But note that the Fourier components of either frequency are themselves badly behaved at infinity on M, which relates to the fact that these are not normalizable states.]

The description of a free photon that has been given above is, in an important sense, a reducible one, since the left- and right-handed photons have been treated together. In fact one can regard these as effectively two distinct particles, their respective wave functions being the anti-self-dual and self-dual parts $F_{ab}^{(-)}$

[2] The fact that we are considering the compactified space M, here, implies that the discussion is more appropriate for massless than massive fields. I am only concerned with massless fields here, but massive ones could also be correspondingly handled. To be "defined on M" a massless field is assumed to have the appropriate Grgin behavior [3] across infinity.

and $F^{(+)}_{ab}$ of F_{ab}:

$$F_{ab} = F^{(-)}_{ab} + F^{(+)}_{ab} \tag{2.2}$$

with

$$(F^{(\pm)}_{ab})^* := \tfrac{1}{2} e_{abcd} F^{(\pm)cd} = \pm i F^{(\pm)ab}$$

e_{abcd} being the alternating tensor. If F_{ab} were real, $F^{(+)}_{ab}$ and $F^{(-)}_{ab}$ would be complex conjugates of one another, but for a wave function they are simply independent complex quantities, each of positive frequency (and therefore necessarily *not* complex conjugates of one another). If $F^{(+)}_{ab} = 0$ we have a left-handed photon (helicity $-\hbar$); if $F^{(-)}_{ab} = 0$ we have a right-handed photon (helicity $+\hbar$).

Equation (2.2) may be reexpressed in the 2-spinor formalism as

$$F_{ab} = \phi_{AB}\, \epsilon_{A'B'} + \epsilon_{AB}\, \tilde{\phi}_{A'B'} \tag{2.3}$$

(the abstract index notation [5] being employed, for convenience), where

$$F^{(-)}_{ab} = \phi_{AB}\, \epsilon_{A'B'} \qquad F^{(+)}_{ab} = \epsilon_{AB}\, \tilde{\phi}_{A'B'}$$

Then for a classical (real) field we would have $\tilde{\phi}_{A'B'} = \overline{\phi}_{A'B'}$ whereas for a photon wave function ϕ_{AB} and $\tilde{\phi}_{A'B'}$ are independent (symmetric) spinors describing the left-handed and right-handed components of the photon [3]. The Maxwell equations (2.1) become

$$\nabla^{AA'} \phi_{AB} = 0 \qquad \nabla^{AA'} \tilde{\phi}_{A'B'} = 0 \tag{2.4}$$

To describe a pure $-\hbar$ helicity state we must take $\tilde{\phi}_{A'B'} = 0$; for a pure $+\hbar$ helicity state we take $\phi_{AB} = 0$.

Let us now turn to the case of gravity. Were we to adopt the common attitude to "gravitons," viz, that they refer to weak-field perturbations of Minkowski space, then we should be led to an analogous description to the above, where two four-index symmetric spinors ϕ_{ABCD}, $\tilde{\phi}_{A'B'C'D'}$ are employed in the Minkowski background space to provide a (complex) linearized curvature tensor

$$K_{abcd} = \phi_{ABCD}\, \epsilon_{A'B'}\, \epsilon_{C'D'} + \epsilon_{AB}\, \epsilon_{CD}\, \tilde{\phi}_{A'B'C'D'} \tag{2.5}$$

The discussion would be exactly as above. However, I am proposing here that we do not work with linear theory. Instead, I wish to apply the electromagnetic analogy to the *full Einstein theory*. Indeed, a decomposition of the full Riemann curvature of a vacuum space-time (with zero cosmological constant) can be expressed [5] in the same form as (2.5):

$$R_{abcd} = \Psi_{ABCD}\, \epsilon_{A'B'}\, \epsilon_{C'D'} + \epsilon_{AB}\, \epsilon_{CD}\, \tilde{\Psi}_{A'B'C'D'} \tag{2.6}$$

with Ψ_{ABCD} and $\tilde{\Psi}_{A'B'C'D'}$ symmetric, where the two terms on the right are the self-dual and anti-self-dual parts of the vacuum curvature, respectively. For a

Nonlinear gravitons and curved twistor theory; The riddle of gravitation – on the occasion of the 60th birthday of Peter G. Bergmann (Proc. Conf., Syracuse Univ., Syracuse, N.Y., 1975). *General Relativity and Gravitation* **7** (1976), 1, 31–52

classicial (real) space-time, $\tilde{\Psi}_{A'B'C'D'}$ must be the complex conjugate of Ψ_{ABCD}. But to describe a (nonlinear) graviton we must expect, by analogy with the electromagnetic case, that Ψ_{ABCD} and $\tilde{\Psi}_{A'B'C'D'}$ will become independent quantities each, in some appropriate sense, of positive frequency. The curvature R_{abcd} thus becomes complex, so we are dealing with some form of *complex* Riemannian space.

Now the concept of a real (pseudo-) Riemannian space can be generalized in more than one way to produce something with a complex curvature. But it should be clear from the electromagnetic analogy how our generalization must proceed. We require a space with a complex symmetric metric g_{ab} (as opposed to, say, a Kähler metric, which would be Hermitian). And since the positive-frequency requirement involves examining the field at *complex* space-time points and demanding that it be holomorphic there, we are led naturally to the idea that the space we are considering should be a complex four-dimensional manifold and that the metric g_{ab} be holomorphic on this manifold. It is easy to generalize the standard general relativity formulas to such situations. The rule is that all quantities that were previously real now become complex; quantities that were previously complex now become "freed" from their complex conjugates, so an expression involving, say, a quantity ξ and its complex conjugate $\bar{\xi}$ is replaced by a similar expression involving two *independent* quantities ξ and $\tilde{\xi}$. Thus, the complex curvature R_{abcd} for a general complex-Riemannian four-manifold has a spinor decomposition [5]

$$\begin{aligned} R_{abcd} = {} & \Psi_{ABCD}\,\epsilon_{A'B'}\,\epsilon_{C'D'} + \epsilon_{AB}\,\epsilon_{CD}\,\tilde{\Psi}_{A'B'C'D'} \\ & + \epsilon_{AB}\,\Phi_{CDA'B'}\,\epsilon_{C'D'} + \epsilon_{CD}\,\Phi_{ABC'D'}\,\epsilon_{A'B'} \\ & + 2\Lambda(\epsilon_{AC}\,\epsilon_{BD}\,\epsilon_{A'C'}\,\epsilon_{B'D'} - \epsilon_{AD}\,\epsilon_{BC}\,\epsilon_{A'D'}\,\epsilon_{B'C'}) \end{aligned} \tag{2.7}$$

where

$$\Psi_{ABCD} = \Psi_{(ABCD)} \quad \text{and} \quad \tilde{\Psi}_{A'B'C'D'} = \Psi_{(A'B'C'D')} \tag{2.8}$$

are independent quantities, and where

$$\Phi_{ABC'D'} = \Phi_{(AB)(C'D')} \quad \text{and} \quad \Lambda \tag{2.9}$$

are complex (satisfying no Hermiticity restrictions—the relations $\tilde{\Phi}_{ABC'D'} = \Phi_{ABC'D'}$ and $\tilde{\Lambda} = \Lambda$ would be redundant equations whose only purpose would be to define the tilde quantities).

For a complex space-time to describe a graviton, we presumably require satisfaction of Einstein's vacuum equations:

$$R_{ab} = 0$$

that is,

$$\Phi_{ABA'B'} = 0 \quad \text{and} \quad \Lambda = 0$$

Then the complex vacuum Bianchi identities become

$$\nabla^{AA'}\Psi_{ABCD} = 0 \qquad \nabla^{AA'}\tilde{\Psi}_{A'B'C'D'} = 0$$

in close analogy with (2.4). But for a pure graviton in a helicity state we also require *either*

$$\tilde{\Psi}_{A'B'C'D'} = 0 \qquad \text{(helicity } -2\hbar\text{)}$$

or

$$\Psi_{ABCD} = 0 \qquad \text{(helicity } +2\hbar\text{)}$$

I shall refer to a complex vacuum space-time for which $\tilde{\Psi}_{A'B'C'D'} = 0$ as *right-flat*; and one for which $\Psi_{ABCD} = 0$ as *left-flat*. Thus, I as proposing that a single left-handed graviton be described by a positive-frequency right-flat (left-curved) complex space-time; and a right-handed graviton by a left-flat space-time (see footnote 3).

It may seem mysterious to use such a description, which is not in any conventionally "physical" sense like an ordinary space-time. But we must bear in mind that a universe whose sole occupant is a single graviton perhaps ought not, in any case, to resemble an ordinary space-time. Furthermore, as with the case of a single photon, we do not now have a field but something which somehow describes the probability distribution for finding the graviton. To make sense of this idea we would need a definition of scalar product between graviton states. Before we could do this, however, it would be necessary to have a reasonable definition of "positive frequency" for a right- or left-flat complex space-time. What is required is some analog of the flat complex space-time $\mathbb{CM}^+$. All this seems a formidable task. It is therefore particularly remarkable and encouraging that the *general* right- (or left-) flat complex (positive frequency?) vacuum space-time can be constructed *automatically* by way of *twistor theory* [1–3] ! There are already various reasons for believing that twistor theory may be able to provide some new insights into the workings of nature. Here we have an apparently independent role for this theory. Previously it had seemed that twistors were only well-suited for the description of phenomena in flat (or conformally flat) space-times. But now it turns out that the most primitive form of space-time curvature (according to the above ideas), namely, a single graviton helicity eigenstate, arises very naturally in twistor theory from the most primitive type of curved twistor space, as we shall see. It will not be possible to go into *full* detail here, but the essential mathematical result will be given.

[3]The concepts of right-flat and left-flat space-time arose originally from E. T. Newman's construction of the spaces [6, 7] referred to as "Heavens." Though much of my discussion has its origins in Newman's work, I shall not have need to refer to it explicitly in this article.

Nonlinear gravitons and curved twistor theory; The riddle of gravitation – on the occasion of the 60th birthday of Peter G. Bergmann (Proc. Conf., Syracuse Univ., Syracuse, N.Y., 1975). *General Relativity and Gravitation* **7** (1976), 1, 31–52

Let us recall [1-3] now complex (compactified) Minkowski space arises from standard (flat) twistor space T. The space T is a complex four-dimensional vector space endowed with a Hermitian form Φ of signature (++--), normally written (see footnote 4)

$$\Phi = Z^0\overline{Z^2} + Z^1\overline{Z^3} + Z^2\overline{Z^0} + Z^3\overline{Z^1}$$
$$= Z^\alpha \overline{Z}_\alpha \qquad (3.1)$$

the (complex) components of the (abstract) twistor Z^α being Z^0, Z^1, Z^2, Z^3, and those of the (abstract) complex conjugate twistor $\overline{Z}_\alpha$ being

$$\overline{Z}_0 = \overline{Z^2} \quad \overline{Z}_1 = \overline{Z^3} \quad \overline{Z}_2 = \overline{Z^0} \quad \overline{Z}_3 = \overline{Z^1}$$

The three portions of T at which Φ is positive, negative, or zero, respectively, are labeled T^+, T^-, and N. Complex two-dimensional linear subspaces of T correspond to the points of CM according to the following scheme:

$$P^{01} : P^{02} : P^{03} : P^{12} : P^{13} : P^{23}$$
$$= \frac{-i}{\sqrt{2}} w_a w^a : w^2 + iw^3 : -w^0 - w^1 : w^0 - w^1 : -w^2 + iw^3 : i\sqrt{2} \qquad (3.2)$$

where

$$P^{\alpha\beta} = X^\alpha Y^\beta - X^\beta Y^\alpha$$

represents the linear span of twistors X^α, Y^α and where w^a is the position vector (see footnote 5) of the point of CM it corresponds to. The linear two-dimensional spaces ("2-spaces") lying in N define the points of *real* (compactified) Minkowski space M; and those lying in $T^+ \cup \{0\}$ define the points of CM^+.

Often it is convenient to consider the projective twistor space PT (where the prefix P means "projective"). The points of CM are then represented in PT by complex projective lines (which are compact holomorphic curves having topology S^2); the points of M are represented by projective lines in PN and those of CM^+ by projective lines in PT^+. Two points of CM have a *null* separation if and only if the corresponding projective lines in PT *meet*.

There is also a special point I, of M, which is the vertex of the light cone at infinity. This corresponds to a special projective line I in PN, described by a skew twistor $I^{\alpha\beta}$ or, equivalently, by its dual $I_{\alpha\beta} = \frac{1}{2}\epsilon_{\alpha\beta\gamma\delta}I^{\gamma\delta}$ (where $\epsilon_{\alpha\beta\gamma\delta} = \epsilon_{[\alpha\beta\gamma\delta]}$; $\epsilon_{0123} = 1$), the only nonzero components being

$$I_{23} = -I_{32} = 1 = I^{01} = -I^{10} \qquad (3.3)$$

[4] The symbol $\overline{Z^2}$ denotes simply the complex conjugate of the complex number Z^2, etc.
[5] The space-time signature is (+ - - -).

The 15-parameter group of linear transformations of T which preserves the form (3.1) and the twistor $\epsilon_{\alpha\beta\gamma\delta}$ (hence unimodular) is SU(2,2). This corresponds to the conformal group action (four-fold-covered) on M. The 10-parameter subgroup of SU(2,2) which preserves $I^{\alpha\beta}$ (or, equivalently, $I_{\alpha\beta}$) as well corresponds to the Poincaré group action on M. Differential forms which play important roles in twistor theory are

$$\iota = I_{\alpha\beta} Z^\alpha dZ^\beta \tag{3.4}$$

$$\tau = \tfrac{1}{2}\, d\iota = \tfrac{1}{2}\, I_{\alpha\beta} dZ^\alpha \wedge dZ^\beta \tag{3.5}$$

(the form τ being degenerate, of rank 2),

$$\rho = \tfrac{1}{6}\, \epsilon_{\alpha\beta\gamma\delta} Z^\alpha dZ^\beta \wedge dZ^\gamma \wedge dZ^\delta \tag{3.6}$$

and

$$\omega = \tfrac{1}{4}\, d\rho = \tfrac{1}{24}\, \epsilon_{\alpha\beta\gamma\delta}\, dZ^\alpha \wedge dZ^\beta \wedge dZ^\gamma \wedge dZ^\delta \tag{3.7}$$

The forms ρ and ω are invariant under the 15-parameter group (i.e., conformally invariant), while ι and τ are invariant only under the 10-parameter group (i.e., Poincaré invariant).

The degenerate 2-form τ, being of rank 2, defines a complex 2-plane element at each point of T. Since $d\tau = 0$ these 2-plane elements are integrable and the whole of T acquires the structure of a (trivial) fiber bundle, the fibers being these integrated 2-planes each with the complex (affine) structure of $\mathbb{C}^2$, and the base space having the complex vector space structure of $\mathbb{C}^2$. In terms of components, the fibers are parametrized by (Z^0, Z^1), for fixed Z^2 and Z^3, where (Z^2, Z^3) parametrizes the base space. The projective space PT also acquires a fiber bundle structure (now no longer trivial) provided we exclude the line I. The fibers are now simply the planes (less I) through the line I. The base space has the topology of S^2 and is a complex projective line ($\mathbb{CP}^1$).

We may think of the form τ as defining a complex area measure on the base space for the fibering of T ("horizontal measure"); and ω defines an area measure μ on each fiber ("vertical measure") via

$$\omega = \tau \wedge \mu \tag{3.8}$$

The additional information given by the forms ι and ρ, over and above that given by τ and ω, is contained in knowledge of the vector field

$$\Upsilon = Z^\alpha \frac{\partial}{\partial Z^\alpha} \tag{3.9}$$

(the "Euler operator"), whose complex integral curves may be regarded as the points of PT. In a standard notation [8] we have

$$\iota = \Upsilon \lrcorner \tau \qquad \rho = \Upsilon \lrcorner \tilde{\omega} \tag{3.10}$$

Nonlinear gravitons and curved twistor theory; The riddle of gravitation – on the occasion of the 60th birthday of Peter G. Bergmann (Proc. Conf., Syracuse Univ., Syracuse, N.Y., 1975). *General Relativity and Gravitation* **7** (1976), 1, 31–52

To construct a nonlinear graviton it will be necessary to replace flat twistor space T by a new space $\mathcal{T}$ whose *local* holomorphic structure is the same as that of T, but which, as far as its large-scale structure is concerned, is somewhat deformed (i.e., "curved"). The local complex manifold structure of $\mathcal{T}$, the vector field T, the forms $\iota, \tau, \rho, \omega$, and the bundle structure of T will all remain (see footnote 6). (The bundle structure is, after all, a consequence of the local properties of the form τ.) However, the precise *linear* structure of T will *not* be retained. This is the crucial point and it will be elaborated shortly. The idea is that similarly to the way in which linear 2-spaces in T represent points in CM, certain analogs of linear two-spaces in $\mathcal{T}$ represent points of a new space $\mathcal{G}$. This four-complex-dimensional space $\mathcal{G}$ possesses a natural complex metric which is right-flat. In fact the *general* right-flat (vacuum) 4-space can be obtained by such a construction. Finally, by the insistance that $\mathcal{T}$ is, in an appropriate sense, analogous to T^+ (or, preferably, to $\mathrm{T}^+ \cup \mathrm{N}$), we can provide a tentative definition of a positive frequency $\mathcal{G}$. Such a space $\mathcal{G}$ describes, according to the ideas presented here, a general (left-handed) graviton.

I shall begin by considering a somewhat simplified situation. I wish to consider the projective space PT in its role simply as a complex manifold. Suppose we make a small but finite deformation of its complex structure (see footnote 7). Now there is a theorem [9, 10] that any complex manifold which is a deformation of PT (i.e., complex projective three-space CP^3) must in fact be identical with PT as a complex manifold. Furthermore, the linear structure must remain as before. Therefore we get nothing new by deforming the complex structure of PT as a whole. However, if we first select a suitable *portion* of PT and deform *its* complex structure then we *can* get something new. Such a suitable portion would be the region PT^+ or a neighborhood $\mathcal{L}$ of some projective line L in PT.

Consider the case of $\mathcal{L}$ for the moment (PT^+ is essentially the same). The projective lines lying in $\mathcal{L}$ correspond, in CM, to the points of a neighborhood of a point L in CM (L being the point of CM corresponding to L). These projective lines are compact holomorphic curves, each having the topology S^2. There will be other compact holomorphic curves in $\mathcal{L}$ (conics, cubics, etc.) which are distinguished from the projective lines by the fact that they "wrap around" more than once in their imbedding in $\mathcal{L}$; that is to say, they belong to different homology classes from that to which the projective lines belong, so they cannot be deformed continuously into single projective lines. I am not concerned with these more complicated holomorphic curves here. But if we consider compact

[6] But not Φ, since it is not holomorphic.

[7] As with a real manifold, a complex manifold can be defined in terms of overlapping coordinate patches, the transition functions in this case being complex-analytic (holomorphic). To deform the complex structure of the manifold, we simply vary these transition functions.

holomorphic curves belonging only to the homology class of the projective lines in $\mathcal{L}$, then we infer from general theory [11] that the only ones are the projective lines themselves. Thus, the local condition of holomorphicity (i.e., merely of differentiability in the complex sense), together with the topological requirements of compactness and correct homology class, are *by themselves* (see footnote 8) sufficient to single out the required four-parameter family–the projective lines–which represent points of CM. The linear structure of PT is not required as an independent quality. It is already implicit in the holomorphic and global properties of $\mathcal{L}$.

Let us now deform the complex structure of $\mathcal{L}$ to obtain a new complex 3-manifold $\mathcal{L}'$. The first question to be asked is whether appropriate analogs of projective lines will still exist. The requirement will be that such lines are to be compact holomorphic curves as before, belonging to the homology class corresponding to that to which the projective lines in $\mathcal{L}$ belong. Now, fortunately, there is a theorem, due to Kodaira [12], which handles just this kind of situation. What must be examined is the sheaf cohomology group $H^1(\mathbf{L}; \Sigma)$, where Σ is the sheaf over $\mathbf{L}$ of germs of holomorphic sections of the normal bundle (see footnote 9) of $\mathbf{L}$ in $\mathcal{L}$. If $H^1(\mathbf{L}, \Sigma) = 0$, then, according to Kodaira's theorem, an appropriate analog of $\mathbf{L}$ will exist in every sufficiently slight deformation $\mathcal{L}'$ of $\mathcal{L}$. This means that if $\mathcal{L}(t)$ is a family of complex manifolds (suitably holomorphic in t) defined for some neighborhood of the origin in the C of the parameter t, where $\mathcal{L}(0) = \mathcal{L}$, then a (holomorphic) family $\mathbf{L}(t) \subset \mathcal{L}(t)$ will exist with $|t| < \rho$, for some sufficiently small $\rho > 0$, for which $\mathbf{L}(0) = \mathbf{L}$. In this sense, as analog of $\mathbf{L}$ will exist in each $\mathcal{L}(t)$ with $|t| < \rho$.

In fact $H^1(\mathbf{L}, \Sigma) = 0$ *is* actually valid. (This is a consequence (see footnote 10) of the fact that the normal bundle of $\mathbf{L}$ in $\mathcal{L}$ is the direct sum of the normal bundles of $\mathbf{L}$ in two planes through $\mathbf{L}$, of Serre duality [13] applied to each of these smaller bundles, and the fact that neighboring lines to $\mathbf{L}$ in these planes meet $\mathbf{L}$.) Thus, provided our choice of $\mathcal{L}'$ lies sufficiently close to $\mathcal{L}$ as regards its complex structure, we are assured that it contains an analog of $\mathbf{L}$. The same applies to each other member of the four-parameter system of projective lines in $\mathcal{L}$. From this it can be seen that $\mathcal{L}'$ (sufficiently like $\mathcal{L}$) contains precisely a four-parameter system of compact holomorphic curves belonging to the correct homology class. Each one of these curves can be used to represent a point in an abstract space $\mathcal{G}$. Thus, we see that $\mathcal{G}$ is four-complex-dimensional, in fact a complex manifold.

The space $\mathcal{G}$ has, naturally defined by the construction, a complex con-

[8] Such behavior is, of course, at complete variance with what occurs in the real case. Differentiable (or analytic) real curves can be varied locally in a great many ways. No finite-dimensional system could ever be defined in this way.

[9] For the meanings of these terms, see references [11, 13].

[10] I am deeply indebted to M. F. Atiyah for supplying this argument that $H^1(\mathbf{L}, \Sigma) = 0$, and also for drawing my attention to Kodaira's paper [12].

Nonlinear gravitons and curved twistor theory; The riddle of gravitation – on the occasion of the 60th birthday of Peter G. Bergmann (Proc. Conf., Syracuse Univ., Syracuse, N.Y., 1975). *General Relativity and Gravitation* **7** (1976), 1, 31–52

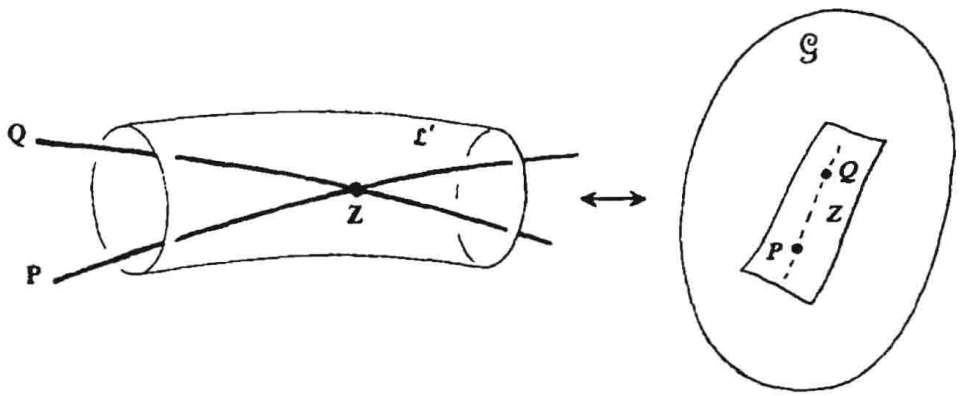


Fig. 1. Points P, Q of $\mathcal{G}$ are null-separated if and only if the corresponding holomorphic curves **P**, **Q** in $\mathcal{L}'$ meet, their intersection point **Z** then representing a totally null plane in $\mathcal{G}$.

formal structure (i.e., that given by a complex metric up to a local proportionality factor). To define this conformal metric it would suffice to give a criterion, for two neighboring points of $\mathcal{G}$, for deciding whether they are null-separated (hence giving a definition of the "null cones" in $\mathcal{G}$). In fact a definition can be given which applies to points which need not be neighboring. If P and Q are two points of $\mathcal{G}$, corresponding to compact holomorphic curves **P** and **Q**, respectively, then we define P and Q to be null-separated if and only if **P** and **Q** have a point in common (see Figure 1).

It needs to be verified that this "conformal structure" is indeed quadratic (i.e., arising from a complex Riemannian metric). We must examine the situation for P and Q neighboring. In effect this means that **Q** can be regarded as a holomorphic cross section of the normal bundle (see footnote 11) of **P** in $\mathcal{L}'$. Those cross sections with zeros define the null cone at P in $\mathcal{G}$. The quadratic nature of this null cone follows from general results concerning deformations of complex structures. For we know that it is quadratic in the original flat case, whereby **CM** is obtained by the same construction from **PT**. (This is just the original twistor construction described in Section 3. In this case **CM** is the "Klein representation" of lines in **PT**; so **CM** is a quadric and its null cones are obtained from tangent hyperplane sections–which are quadric cones, i.e., quadratic.) When we deform the structure to obtain $\mathcal{L}'$, these null cones are deformed. The system of null rays at a point of $\mathcal{G}$ (generators of the null cone) constitutes a compact complex 2-manifold $\mathcal{C}$ which is a quadric in the flat case, and which in general is obtained from that of the flat situation by deforming the complex structure of this quadric. But it is a theorem [14] that every deformation of the complex structure of a quadric is trivial (i.e., gives us simply the quadric back again). Thus the conformal structure of $\mathcal{C}$ *is* quadratic.

[11] See references [11, 13] for the definition of these terms.

Armed with this information we can now proceed to investigate the curvature of $\mathcal{G}$. For any choice of metric compatible with its conformal structure we shall have a spinor decomposition (2.7). The Ricci parts $\Phi_{ABC'D'}$ and Λ of the curvature are not conformally invariant, so little can be said about them. However, each of the spinors Ψ_{ABCD} and $\tilde{\Psi}_{A'B'C'D'}$ *is* conformally invariant, corresponding to the anti-self-dual and self-dual parts, respectively, of the Weyl conformal tensor. The identification of these two 4-valent spinors in terms of $\mathcal{L}'$ depends upon identifying the "unprimed" and "primed" spin-spaces at a point P of $\mathcal{G}$. For this, we must examine the null cone at P and see how its space $\mathcal{C}$ of generators (null rays through P) may be expressed as a product of two $\mathbb{CP}^1$'s– the unprimed and primed projective spin-spaces at P.

A complex null vector V^a at P will have spinor expression

$$V^a = \xi^A \eta^{A'} \tag{4.1}$$

(using abstract indices [5]). Holding ξ^A fixed and varying $\eta^{A'}$ we obtain a linear 2-space of null vectors V^a, this 2-space being defined by ξ^A up to proportionality; holding $\eta^{A'}$ fixed and varying ξ^A we obtain a linear 2-space, defined by $\eta^{A'}$ up to proportionality. The family of 2-spaces defined by the various ξ^A's up to proportionality gives us the *unprimed projective spin-space* at P; the family defined by the various $\eta^{A'}$'s up to proportionality gives us the *primed projective spin-space* at P. Using the standard twistor conventions [1], [3], we find that, in the case of flat twistor space, the various *planes* through the line $\mathbf{P}$ in $\mathcal{L}$ correspond to the various unprimed projective spinors at P, whereas *points* on the line P correspond to primed projective spinors at P. That is to say, if we fix a plane ξ through $\mathbf{P}$, then the lines in ξ neighboring $\mathbf{P}$ correspond to a family of vectors (4.1) for which ξ^A is held fixed and $\eta^{A'}$ varied; if we fix a point η on $\mathbf{P}$, then the lines through η neighboring $\mathbf{P}$ correspond to a family of vectors (4.1) for which $\eta^{A'}$ is held fixed and ξ^A varied. For a "curved" $\mathcal{L}'$, the situation is similar except that instead of having analogs of complete planes in $\mathcal{L}'$ we have only a plane "strip" through $\mathbf{P}$ of infinitesimal width, corresponding to each unprimed projective spinor at P. The points of $\mathbf{P}$, as before, correspond to the primed projective spinors at P.

Now the crucial feature of the above description is that whereas in $\mathcal{L}$ we have both points and planes, in $\mathcal{L}'$ we normally have only points, the plane strips being generally not integrable to complete planes. To interpret this feature in $\mathcal{G}$, we examine the system of lines (see footnote 12) in $\mathcal{L}'$ through a point $\mathbf{Z}$ of $\mathcal{L}'$. This system will be two-complex-dimensional (since passing through a point reduces the dimensionality of the line system by 2). Thus the system corresponds to a complex 2-surface Z in $\mathcal{G}$. At a point Q of Z (corresponding to a line $\mathbf{Q}$ through $\mathbf{Z}$ in $\mathcal{L}'$) the tangent plane to Z will be given by vectors of the form

[12] Henceforth I shall use the term "line" to denote a compact holomorphic curve in $\mathcal{L}'$ belonging to the appropriate homology class.

(4.1) for fixed $\eta^{A'}$ and varying ξ^A. Note that all such vectors are orthogonal to one another (since the $\eta^{A'}$'s annihilate each other by the skewness of ϵ_{AB}). Thus, Z is a *totally null* surface. Every tangent vector to Z will have the form (4.1) for some ξ^A, where $\eta^{A'}$ can be taken to be a fixed spinor field on Z. Now, the integrability condition for this (i.e., the condition that the Lie bracket of any two of these vectors again has the same form) gives, upon transvection with $\eta_{A'}$, the relation (see footnote 13)

$$\eta_{A'}\eta^{B'}\nabla_{BB'}\eta^{A'} = 0 \tag{4.2}$$

∇_a being a Riemannian covariant derivative for some metric compatible with the conformal structure of $\mathcal{G}$. Operating on (4.2) with $\eta^{C'}\nabla^{B}_{C'}$ (with contraction on B) we obtain, after some simplification and use of the spinor Ricci identities [5] the relation

$$\tilde{\Psi}_{A'B'C'D'}\eta^{A'}\eta^{B'}\eta^{C'}\eta^{D'} = 0 \tag{4.3}$$

At each point of $\mathcal{G}$, (4.3) must hold for *every* $\eta^{A'}$ (since, corresponding to each line of $\mathcal{L}'$ we have the above situation for every one of its points). Hence (invoking the symmetry of $\tilde{\Psi}_{A'B'C'D'}$) we derive

$$\tilde{\Psi}_{A'B'C'D'} = 0 \tag{4.4}$$

at every point of $\mathcal{G}$. Thus, $\mathcal{G}$ is automatically *conformally right-flat*. But $\mathcal{G}$ is generally not conformally left-flat because planes do not exist in $\mathcal{L}'$.

It is not hard to reverse the above argument and derive also the converse result that *every* conformally right-flat 4-space is obtainable by the above construction. Of course a corresponding construction would apply to conformally *left-flat* spaces, but conventionally I shall say that these are obtained by deforming the structure of *dual* projective twistor space PT*.

§(5): *Curved Twistor Space*

To proceed further we need the concept of a metric g_{ab} on $\mathcal{G}$ and its associated Riemannian connection ∇_a. For these it is convenient to consider deformations of (portions of) *nonprojective* twistor space T, since the forms ω and τ will be required (and these are not defined on PT). The deformed (appropriate) portion of T will be denoted by $\mathcal{T}$, where P$\mathcal{T}$ is the space $\mathcal{L}'$ (or PT$^{+\prime}$) considered above. The passage from $\mathcal{T}$ to P$\mathcal{T}$ is achieved by factoring out by the integral curves of the (holomorphic) vector field Υ, this being the deformed version of (3.9). The form τ is to be an exact degenerate 2-form satisfying

$$\tau = \tfrac{1}{2}\, d\iota \tag{5.1}$$

[13] One consequence of (4.2) is that Z is a totally geodesic surface. Consequently the null-separation depicted in Figure 1 really means null-geodesic-separation.

where

$$\iota = \Upsilon \lrcorner \tau \tag{5.2}$$

just as before [cf. (3.5), (3.10)]. Similarly, the 4-form ω satisfies

$$\omega = \tfrac{1}{4}\, d\rho \tag{5.3}$$

where

$$\rho = \Upsilon \lrcorner \omega \tag{5.4}$$

as before [cf. (3.7), (3.10)]. Again the degeneracy and closedness of τ implies that it defines a 2-foliation; I assume that, as before, this is a fibering of $\mathcal{J}$ by 2-surfaces. This fibering is invariant under the action of Υ since from (5.1) and (5.2) we obtain [8]

$$\underset{\Upsilon}{\pounds}\, \tau = 2\tau \tag{5.5}$$

Similarly (5.3) and (5.4) imply

$$\underset{\Upsilon}{\pounds}\, \omega = 4\omega \tag{5.6}$$

so interpreting Υ as a homogeneity (Euler) operator we can say that τ and ω are forms of homogeneity degrees 2 and 4, respectively.

The base space of the fibering (see footnote 14) is to have the complex structure

$$\mathbb{C}^2 - \{0\} \tag{5.7}$$

and will be interpreted, in effect, as the *primed spin-space* (less the zero spinor) for $\mathcal{G}$. The essential significance of this fibering and interpretation arises from the following. We intend to produce a space $\mathcal{G}$ which is right-flat. This means that the metric of $\mathcal{G}$ that we construct must satisfy

$$\Psi_{A'B'C'D'} = 0 \qquad \Phi_{ABA'B'} = 0 \qquad \Lambda = 0 \tag{5.8}$$

Examination of the spinor Ricci identities [5] yields the conclusion that (5.8) is necessary and sufficient for covariant derivatives to commute when applied to primed spinors

$$(\nabla_a \nabla_b - \nabla_b \nabla_a)\, \eta^{C'} = 0 \tag{5.9}$$

from which it follows that there is then an integrable (global) parallelism of covariantly constant primed spinors at points of $\mathcal{G}$. The space of the different possible (nonzero) constant primed spinors will be identifiable with (5.7).

[14] The vector Υ projects to a vector field on $\mathbb{C}^2 - \{0\}$ which is required to be the homogeneity (Euler) operator on $\mathbb{C}^2$.

Nonlinear gravitons and curved twistor theory; The riddle of gravitation – on the occasion of the 60th birthday of Peter G. Bergmann (Proc. Conf., Syracuse Univ., Syracuse, N.Y., 1975). *General Relativity and Gravitation* **7** (1976), 1, 31–52

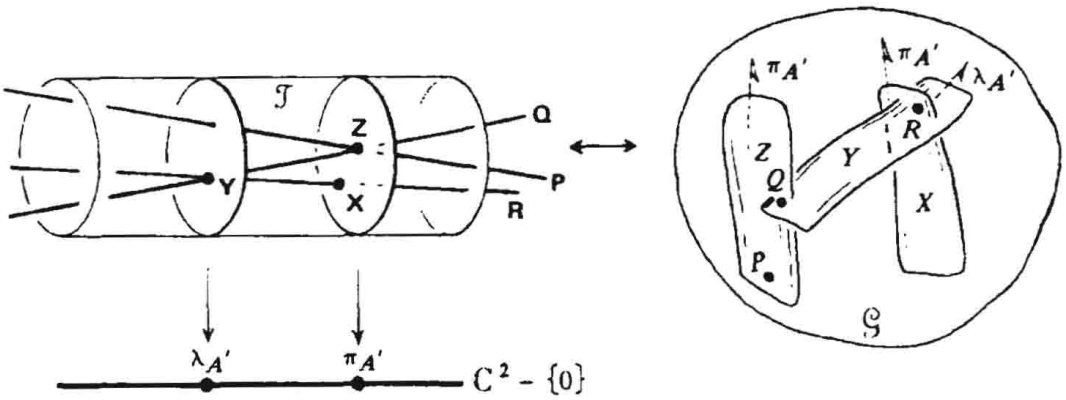


Fig. 2. Points of $\mathcal{G}$ are represented by holomorphic cross sections of the bundle $\mathcal{T}$. The global parallelism for primed spinors in $\mathcal{G}$ gives rise to the bundle structure of $\mathcal{T}$.

Our construction (see Figure 2) will identify points of $\mathcal{G}$ with *holomorphic cross sections* of $\mathcal{T}$ regarded as a fiber bundle over $\mathbb{C}^2 - \{0\}$, the sections being tangent to the vector field Υ. Thus, if we factor out $\mathcal{T}$ by the integral curves of Υ to obtain $\mathbb{P}\mathcal{T}$, these cross sections become compact holomorphic curves (belonging necessarily to the appropriate homology class) in agreement with the construction given before. Now choose a point **Z** in $\mathcal{T}$ and consider the various cross sections **P** through **Z**. The corresponding points P of $\mathcal{G}$ describe a totally null surface with tangent vectors of the form $\xi^A \pi^{A'}$, for varying ξ^A, where we can scale (see footnote 15) the fixed $\pi_{A'}$ to be *covariantly constant* over Z:

$$\pi^{B'} \nabla_{BB'} \pi_{A'} = 0 \tag{5.10}$$

[The integrability condition for achieving (5.10), with $\pi_{A'}$ a suitable multiple of the $\eta_{A'}$ of (4.2), is $\Psi_{A'B'C'D'} \pi^{A'} \pi^{B'} \pi^{C'} = 0$. This is satisfied because of (4.4).] Given the surface Z, the freedom in the choice of $\pi^{A'}$ is just an overall constant scaling. This freedom corresponds to the fact that different choices of "twistor" $\mathbf{Z} \in \mathcal{T}$ can be made which are associated with the same "projective twistor" $\mathbf{Z} \in \mathbb{P}\mathcal{T}$. The twistor **Z** itself is to encompass both the information of the location of the surface Z in $\mathcal{G}$ and of the scaling for $\pi_{A'}$.

Now holding **Z** fixed and varying the holomorphic cross section **P** through it, we get, in $\mathcal{T}$, a system of points with *constant* $\pi_{A'}$. If we vary **P** so that it no longer passes through **Z** we get other points in $\mathcal{G}$. The global parallelism for $\pi_{A'}$-spinors extends also to these points, so there will be a family of different Z's all of which correspond to the *same* constant $\pi_{A'}$. Each such family of Z's will constitute a *fiber* in the bundle structure of $\mathcal{T}$. Thus, the fibers are labeled by the different $\pi_{A'}$'s, the base-space (5.7) being the *space* of possible $\pi_{A'}$'s.

The above discussion shows, in essence, that *if* we are to obtain a right-flat space by such a construction, then the structure of $\mathcal{T}$ must be essentially as

[15] The choice of the "lowered" version of $\pi_{A'}$ in (5.10) is for reasons of conformal invariance. This choice fits in with normal twistor conventions [1, 3].

described. To proceed in the other direction is more difficult. First the metric of $\mathcal{G}$ must be identified in terms of the structure of $\mathcal{T}$, then the connection must be identified, and finally the right-flat nature of the curvature must be established. I shall give the argument partly in outline.

The identification of the *metric* in terms of the forms τ, ω is not quite so direct as one might have hoped. Roughly speaking it is the following. Let **P** be the appropriate holomorphic cross section of $\mathcal{T}$ corresponding to the point P of $\mathcal{G}$ and let **Q** and **R** be the neighboring cross sections to **P** corresponding to two points Q and R neighboring to P (i.e., in the tangent space to P in $\mathcal{G}$). Label these tangent vectors (i.e., $\overrightarrow{PQ}, \overrightarrow{PR}$) by U^a, V^a, respectively. What we require is

$$g(U, V) = g_{ab}U^aV^b \tag{5.11}$$

Now **Q** and **R** may be regarded as holomorphic cross sections of the normal bundle of **P** in $\mathcal{T}$. Because of the nature of the fibering of $\mathcal{T}$, the fibers of the normal bundle of **P** can be identified as the various tangent 2-spaces at the intersections of **P** with the fibers of the bundle $\mathcal{T}$. Since each fiber $\mathcal{F}_\pi$ of $\mathcal{T}$ has a symplectic structure (area element) μ such that, as in (3.8),

$$\omega = \tau \wedge \mu \tag{5.12}$$

it follows that there is a *skew* "scalar product" $\delta_\pi(U, V)$ for each fiber of the normal bundle. (Intuitively, we examine where **P** and the neighboring cross sections **Q** and **R** intersect $\mathcal{F}_\pi$. We draw vectors from the intersection of **P** with $\mathcal{F}_\pi$ to the respective intersections of **Q** and **R** with $\mathcal{F}_\pi$, and evaluate 2μ on this pair of vectors to obtain $\delta_\pi(U, V)$.

Now as $\mathcal{F}_\pi$ varies (i.e., as the primed spinor $\pi_{A'}$ varies over the base space of $\mathcal{T}$) we obtain a function of $\pi_{A'}$ which is holomorphic and homogeneous of degree 2 [since, by equations (5.5) and (5.6), μ is homogeneous of degree 2 under Υ]. The only such functions of $\pi_{A'}$ (defined on the whole of $\mathbb{C}^2 - \{0\}$) are quadratic. The coefficients of this quadratic form serve to define a quantity which I write

$$U^{A(A'}V^{B')}_{A} \tag{5.13}$$

giving

$$\delta_\pi(U, V) = U^{AA'}V^{B'}_{A}\pi_{A'}\pi_{B'} \tag{5.14}$$

We may regard (5.13) or (5.14) as providing, in effect, the information of one-half of the metric g_{ab}, namely, the ϵ_{AB} half in the abstract index relation [5]

$$g_{ab} = \epsilon_{AB}\epsilon_{A'B'} \tag{5.15}$$

(since ϵ_{AB} is involved in the lowering of the index A in $V^{AB'}$). The remaining information, $\epsilon_{A'B'}$ (or rather its inverse $\epsilon^{A'B'}$), is defined by the form τ, since τ provides a complex simplectic structure (area element) on $\pi_{A'}$-space.

Nonlinear gravitons and curved twistor theory; The riddle of gravitation – on the occasion of the 60th birthday of Peter G. Bergmann (Proc. Conf., Syracuse Univ., Syracuse, N.Y., 1975). *General Relativity and Gravitation* **7** (1976), 1, 31–52

The obtaining of (5.11) explicitly from (5.14) is a little indirect—not surprisingly, perhaps, since (5.11) is symmetric in U, V whereas (5.14) is skew. It is easier if we restrict attention to the case when U^a and V^a are *null* vectors. The definition of (5.11) in the general case will then follow by linearity. This requirement of nullness means that **P** must now intersect each of **Q** and **R**. Let an intersection (see footnote 16) of **P** with **Q** lie in the fiber $\mathcal{F}_\alpha$ and an intersection of **P** with **R** lie in $\mathcal{F}_\beta$ (i.e., respectively, above $\alpha_{A'}$ and $\beta_{A'}$ in $\pi_{A'}$-space). Then we have

$$g(U, V) = \frac{(\alpha^{C'}\beta_{C'})}{(\alpha^{A'}\pi_{A'})(\beta^{B'}\pi_{B'})}\,\delta_\pi(U, V) \tag{5.16}$$

for the definition of the metric in terms of $\delta_\pi(U, V)$. (The spinor scalar products $\alpha^{C'}\beta_{C'}$, etc. refer to the linear structure on $\mathbb{C}^2 - \{0\}$ and the complex measure on it given by τ. Explicit coordinate expressions will be given in Section 6.) The fact that (5.16) is actually independent of $\pi_{A'}$ follows from the fact that it is homogeneous of degree zero and defined holomorphically over the whole of $\mathbb{C}^2 - \{0\}$.

Next I shall show how the *connection* on $\mathcal{G}$ may be defined in relation to $\mathcal{T}$. It is sufficient to interpret, in $\mathcal{G}$, the parallel propagation of an arbitrary vector U^a along a curve γ which starts from an arbitrary point of $\mathcal{G}$ in an arbitrary direction. Indeed, by linearity, it will be enough to consider only the case when U^a is a null vector and the curve γ is also null. I shall suppose, in fact, that γ lies in the totally null plane Z, in $\mathcal{G}$, corresponding to the cross sections of $\mathcal{T}$ which pass through a fixed point **Z** of $\mathcal{T}$. With notation as before, we have a fiber $\mathcal{F}_\alpha$ corresponding to the primed-spinor part of U^a:

$$U^a = \kappa^A \alpha^{A'} \tag{5.17}$$

Now the vector U^a at the point P (on γ) is determined, in $\mathcal{T}$, by knowledge of **P**, of $\mathcal{F}_\alpha$, and of the vector q in $\mathcal{F}_\pi(\ni Z)$ which joins the intersection of **P** with $\mathcal{F}_\pi$ to that of **Q** with $\mathcal{F}_\pi$. (Obtain **Q** by joining the "tip" of the vector q to the point $\mathbf{P} \cap \mathcal{F}_\alpha$.) To propagate U^a along γ (in $\mathcal{G}$) we simply hold q and $\mathcal{F}_\alpha$ *fixed* as we vary **P** (corresponding to moving P along γ). See Figure 3. Note that this parallel propagation is integrable provided γ remains in a fixed Z. But if we propagate U^a in directions corresponding to different Z's and to different fibers $\mathcal{F}_\pi$, then q will generally change in a nonintegrable way. Nevertheless $\mathcal{F}_\alpha$ will remain unchanged however P moves. From this we infer that the κ^A of (5.17) moves nonintegrably under parallel propagation, in general, whereas $\alpha^{A'}$ achieves an integrable parallelism throughout $\mathcal{G}$, this parallelism being achieved by holding $\alpha^{A'}$ in the base space fixed.

[16] The intersection would be unique in $P\mathcal{T}$, but here we have an intersection above each $\lambda\alpha_{A'}$, with λ a nonzero complex number. Note that (5.16) is invariant under each rescaling of $\alpha_{A'}$ and $\beta_{A'}$.

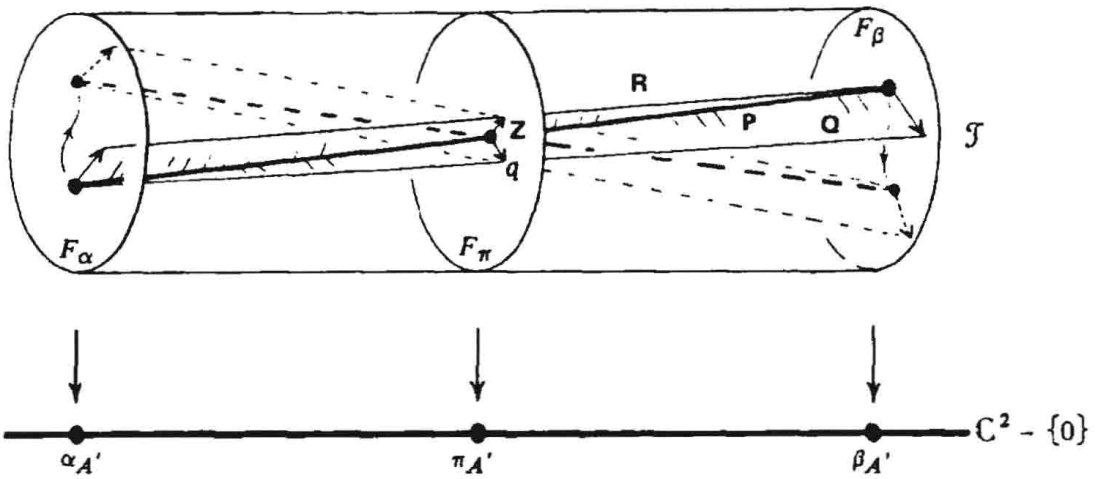


Fig. 3. The parallel propagation, in $\mathcal{G}$, of null vectors $\overrightarrow{PQ}$ and $\overrightarrow{PR}$ is depicted in terms of $\mathcal{T}$. The propagation direction is in the 2-plane in $\mathcal{G}$ defined by **Z**. The constancy of $\delta_\pi(U, V)$ is established by examining three distinct $\pi_{A'}$'s: $\alpha_{A'}, \beta_{A'}, \pi_{A'}$.

To proceed further we must verify that this concept of parallel propagation indeed defines a connection which has vanishing torsion and that the metric is covariantly constant under this connection. I shall omit the first part of the argument (which contains no essential difficulty but is awkward to express) and concentrate only on the constancy of the metric. Note that $g(U, V)$ in (5.16) differs from $\delta_\pi(U, V)$ only by an expression built from primed spinor contractions. The parallel propagation, as defined above, leaves these terms unchanged, so we need examine only its effect on $\delta_\pi(U, V)$. Now we have seen in (5.14) that $\delta_\pi(U, V)$ is quadratic in $\pi_{A'}$. Thus it is determined completely by its values for three distinct choices of $\pi_{A'}$. Select $\alpha_{A'}, \beta_{A'}$, and the value of $\pi_{A'}$ corresponding to the (null) propagation direction as these three choices (see Figure 3) (assuming, as before, that U^a and V^a are also null). We see that $\delta_\alpha(U, V) = 0 = \delta_\beta(U, V)$ throughout the motion, since one of the vectors that μ is to be evaluated on vanishes. Furthermore, for $\pi_{A'}$ corresponding to the propagation direction, $\delta_\pi(U, V)$ does not change according to the definition given. Hence $\delta_\pi(U, V)$ is indeed constant under propagation and so $g(U, V)$ is constant also, as required.

This shows that this concept of propagation is the one given by the Riemannian connection. We have seen that it induces a global parallelism for primed spinors. Thus (5.9) holds, so (5.8) holds also and $\mathcal{G}$ is right-flat.

Finally, in order that $\mathcal{G}$ should qualify as a graviton state, it is necessary to specify the requirement that it should have *positive frequency*. What we require is that, in an appropriate sense, $\mathcal{T}$ contain an analog of $\mathrm{T}^+ \cup \mathrm{N}$, since this implies that the metric and curvature of $\mathcal{G}$ are holomorphic on an appropriately defined analog of CM^+. It is not absolutely clear what this sense should be. But, tenta-

Nonlinear gravitons and curved twistor theory; The riddle of gravitation – on the occasion of the 60th birthday of Peter G. Bergmann (Proc. Conf., Syracuse Univ., Syracuse, N.Y., 1975). *General Relativity and Gravitation* **7** (1976), 1, 31–52

tively, I would suggest that each fiber $\mathcal{F}_\pi$ should be identical in its intrinsic structure with what would be the case if $\mathcal{T}$ were an open portion of flat twistor space containing $\mathcal{T}^+ \cup \mathcal{N}$. That is, with an open portion of $\mathbb{C}^2$ which contains the half-space $\mathrm{Re}(z) \geqslant 0$, where $(z, w) \in \mathbb{C}^2$. The analog of $\mathcal{N} \cap \mathcal{F}_\pi$ would be given, in these coordinates (see footnote 17) by $\mathrm{Re}(z) = 0$ and the analog of $\mathcal{T}^+ \cap \mathcal{F}_\pi$ by $\mathrm{Re}(z) \geqslant 0$.

If we adjoin to $\mathcal{T}$ some extra points suitably defined for when $\pi_{A'} = 0$, then we can arrange that the analog of $\mathcal{PN}$ is ruled by a 3-real-parameter family of compact holomorphic curves. These correspond to the "real" points at *infinity* in $\mathcal{G}$. But generally $\mathcal{T}$ contains no other compact holomorphic curves which could correspond to "real" *finite* points of $\mathcal{G}$. Thus $\mathcal{G}$ is a truly complex space which should not be regarded as the "complexification" of some real space-time.

§(6): *Explicit Determination of Right-Flat Spaces*

In this section I shall show how to specify the structure of a general curved twistor space $\mathcal{T}$ explicitly in terms of a single complex function of three variables. This will give us the general right-flat 4-space. As far as can be seen at the moment, however, the method I am proposing does not appear to give a simple way of ensuring the positive-frequency condition, so I shall not concern myself with this aspect of the construction of nonlinear gravitons here. It may be mentioned that in the (apparently) completely different approach to the construction of right-flat spaces ("Heavens") due to E. T. Newman [6, 7], one can also specify such spaces by means of arbitrary complex functions of three variables (complex "news functions"), and that in his approach there appears to be a way of restricting these functions to obtain what might be plausibly called positive-frequency right-flat spaces. The present approach does have some advantages, however, in its explicit simplicity. There are no explicit differential equations which need to be solved.

The space $\mathcal{T}$ will be built from just two overlapping open sets $\mathcal{Q}$ and $\hat{\mathcal{Q}}$ ($\mathcal{Q} \cup \hat{\mathcal{Q}} = \mathcal{T}$) on which coordinates (z^0, z^1, z^2, z^3) and $(\hat{z}^0, \hat{z}^1, \hat{z}^2, \hat{z}^3)$, respectively, are defined. The transition functions on $\mathcal{Q} \cap \hat{\mathcal{Q}}$ are to be specified in terms of a single holomorphic function

$$f(\hat{z}^0, z^1, z^2, z^3) \tag{6.1}$$

homogeneous (see footnote 18) of degree 2:

$$f(\lambda\hat{z}^0, \lambda z^1, \lambda z^2, \lambda z^3) = \lambda^2 f(\hat{z}^0, z^1, z^2, z^3) \tag{6.2}$$

[17] In the standard coordinates for T given in Section 3 we have $\mathcal{F}_\pi$ defined by $Z^2 = \bar{a}$, $Z^3 = \bar{b}$ (a, b constant). Then we can take $z = aZ^0 + bZ^1$, $w = Z^1$ (assuming $a \neq 0$).

[18] Thus f, in essence, is a holomorphic function of three complex variables.

according to

$$\frac{\partial f}{\partial z^1} = z^0 \qquad \hat{z}^1 = \frac{\partial f}{\partial \hat{z}^0} \qquad \hat{z}^2 = z^2 \qquad \hat{z}^3 = z^3 \tag{6.3}$$

on $\mathfrak{Q} \cap \hat{\mathfrak{Q}}$. The base space of $\mathcal{T}$ is coordinatized by (z^2, z^3) and by $(\hat{z}^2, \hat{z}^3)$. We may regard these as the component of $\pi_{A'}$:

$$\pi_{0'} = z^2 = \hat{z}^2 \qquad \pi_{1'} = z^3 = \hat{z}^3 \tag{6.4}$$

The ratios $\pi_{0'} : \pi_{1'}$ define points of a sphere S^2. We define the sets $\mathfrak{Q}$ and $\hat{\mathfrak{Q}}$ to correspond to a covering of S^2 by two overlapping open cells $\mathfrak{D}$ and $\hat{\mathfrak{D}}$ ($\mathfrak{D} \cup \hat{\mathfrak{D}} = S^2$), the intersection being an annular region $\mathfrak{A} = \mathfrak{D} \cap \hat{\mathfrak{D}}$. The region $\mathfrak{Q}$ is defined to be the set of points of $\mathcal{T}$ for which $z^2 : z^3 \in \mathfrak{D}$, and similarly $\hat{\mathfrak{Q}}$ is the set for which $\hat{z}^2 : \hat{z}^3 \in \hat{\mathfrak{D}}$. (See Figure 4.)

The Euler operator Υ is to have its standard expression [cf. (3.9)] in both coordinate systems. The differential forms

$$\tau = dz^2 \wedge dz^3 = d\hat{z}^2 \wedge d\hat{z}^3 \tag{6.5}$$

and

$$\mu = dz^0 \wedge dz^1 = d\hat{z}^0 \wedge d\hat{z}^1 \tag{6.6}$$

also have their normal expressions. The reason for the particular form of coordinate transformation given by (6.3) is that it preserves Υ and the fibering, leaving the coordinates on the base space unchanged, while the symplectic form μ is

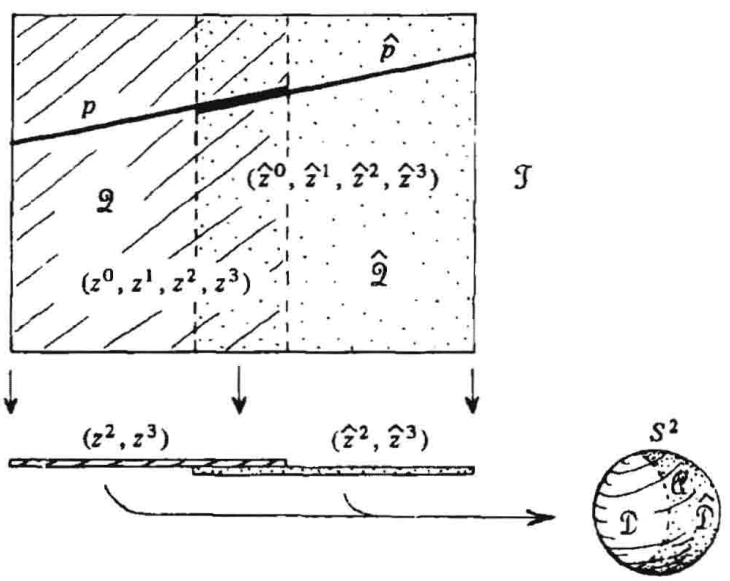


Fig. 4. Two coordinate patches cover $\mathcal{T}$, with overlap in an annular type of region.

preserved on each fiber because the generating function f yields, by (6.3), a canonical transformation on the fibers.

To construct a *point* of $\mathcal{G}$ we need a holomorphic cross section of $\mathcal{J}$. This is achieved by specifying z^0, z^1 as functions of z^2, z^3 in $\mathcal{Q}$ and $\hat{z}^0, \hat{z}^1$ as functions of $\hat{z}^2, \hat{z}^3$ in $\hat{\mathcal{Q}}$:

$$\begin{aligned} z^0 &= p^0(z^2, z^3) \qquad z^1 = p^1(z^2, z^3) \qquad (z^2 : z^3 \in \mathcal{D}) \\ \hat{z}^0 &= \hat{p}^0(\hat{z}^2, \hat{z}^3) \qquad \hat{z}^1 = \hat{p}^1(\hat{z}^2, \hat{z}^3) \qquad (\hat{z}^2 : \hat{z}^3 \in \hat{\mathcal{D}}) \end{aligned} \tag{6.7}$$

where the coordinate transformation (6.3) is to hold on the overlap region $z^2 : z^3 = \hat{z}^2 : \hat{z}^3 \in \mathcal{R}$. Each of $p^0, p^1, \hat{p}^0, \hat{p}^1$ is to be homogeneous of degree 1 in its arguments and holomorphic over the whole range indicated. Such sets of functions [provided (6.3) is not too far from the identity transformation] will form a 4-complex-parameter system, which defines the space $\mathcal{G}$.

As stated, the problem of constructing the points of $\mathcal{G}$ does not explicitly appear to involve the solution of differential equations. However, this is somewhat misleading because a complicated implicit boundary value problem is involved. (If desired, the expressions could be written out in terms of their real and imaginary parts and reformulated in terms of elliptic differential equations.) It is not out of the question that some way of generating the p's explicitly along with the structure of $\mathcal{J}$ might be possible, but no method for doing this has emerged as yet.

Finally, the metric on $\mathcal{G}$ may be determined from the scalar product $g(U, V)$ of two null vectors U^a, V^a at a point P of $\mathcal{G}$, this scalar product being calculated by means of (5.16). If we define P according to (6.7), points Q and R infinitesimally separated from P by displacements ϵU^a and ϵV^a, respectively, can be given by replacing each p-function in (6.7) by $p + \epsilon u$ and $p + \epsilon v$ in turn, where ϵ is infinitesimal and where u stands for a collection of four functions $u^0(z^2, z^3)$, $\ldots$, $\hat{u}^1(\hat{z}^2, \hat{z}^3)$ and similarly for v (and p). Since U^a is null, P and Q will intersect, say for $z^2 = \alpha z^3$ (or $\hat{z}^2 = \alpha \hat{z}^3$). This means that $u^0 = 0 = u^1$ at $z^2 = \alpha z^3$ (or $\hat{u}^0 = 0 = \hat{u}^1$ at $\hat{z}^2 = \alpha \hat{z}^3$, as appropriate). It follows that $(z^2 - \alpha z^3)^{-1} u^0$ and $(z^2 - \alpha z^3)^{-1} u^1$ are nonsingular (as are their careted versions). The same applies to v at some other value of $z^2 : z^3$, say at $z^2 = \beta z^3$. Then we have (see footnote 19)

$$g(U, V) = \frac{(\beta - \alpha)(u^0 v^1 - u^1 v^0)}{(z^2 - \alpha z^3)(z^2 - \beta z^3)} \tag{6.8}$$

by (5.16). The same formula holds in its careted version since 2μ evaluated on $(u^0, u^1), (v^0, v^1)$ is $u^0 v^1 - u^1 v^0$, by (6.6), so

$$\hat{u}^0 \hat{v}^1 - \hat{u}^1 \hat{v}^0 = u^0 v^1 - u^1 v^0$$

[19] The spinor scalar product $\alpha^{A'} \beta_{A'} = \alpha_{1'} \beta_{0'} - \alpha_{0'} \beta_{1'}$ reduces to $\beta - \alpha$ when we adopt $\alpha_{0'} = \alpha$, $\alpha_{1'} = 1$ and $\beta_{0'} = \beta, \beta_{1'} = 1$ to obtain $\alpha^{A'} \pi_{A'} = z^2 - \alpha z^3$ and $\beta^{A'} \pi_{A'} = z^2 - \beta z^3$.

follows from the vector transformation laws for u, v and the canonical nature of (6.3). The relation (6.8) is homogeneous of degree zero and holomorphic on the whole S^2 (of $z^2 : z^3$) and consequently it is independent of z^2, z^3 as required.

§(7): *Outlook*

The above ideas can constitute only the barest beginnings for a new approach to the quantization of general relativity. The positive-frequency requirement would seem to need further clarification. Some definition of normalization and of matrix elements between graviton states is wanted. Also, we need to know how to describe states involving more than one graviton. It seems unlikely that any too immediate concept of Fock space will suffice (see footnote 20) which does not take into account the mutual nonlinearities between the different gravitons. It is more probable that a deformed product of twistor spaces—or some such—will be required. Also we shall need to introduce states of particles for the various matter fields. Here the deformed twistor spaces for the graviton states will be at hand on which to define twistor functions generating these fields [1, 3]. But clearly much more work needs to be done.

Acknowledgments

My thanks are especially due to E. T. Newman and M. F. Atiyah for supplying vital ideas.

References

1. Isham, C. J., Penrose, R., and Sciama, D. W., eds. (1975). *Quantum Gravity: An Oxford Symposium* (Oxford University Press, London).
2. Penrose, R. (1967). *J. Math. Phys.*, **8**, 345.
3. Penrose, R., and MacCallum, M. A. H. (1973). *Phys. Rep.*, **6C**, 242.
4. Streater, R. F., and Wightman, A. S. (1964). *PCT, Spin and Statistics, and All That* (W. A. Bengamin, Inc., New York).
5. Penrose R. (1968). in *Battelle Rencontres 1967* (eds. DeWitt, C. M., and Wheeler, J. A.) (Benjamin, New York).
6. Newman, E. T. report to Tel Aviv Conference 1974 (unpublished).
7. Newman, E. T. (1976). *Gen. Rel. Grav.*, **7**, 107–111.
8. Nelson, E. (1967). *Tensor Analysis* (Princeton University Press, Princeton, New Jersey).
9. Kodaira, K., and Spencer, D. C. (1958). *Ann. Math.*, **67**, 328, 403.
10. Hirzebruch, F., and K. Kodaira, K. (1957). *J. Math. Pures Appl.*, **36**, 201.
11. Hirzebruch, F. (1962). *New Topological Methods in Algebraic Geometry* (Springer, Berlin).
12. Kodaira, K. (1963). *Amer. J. Math.*, **85**, 79.
13. Gunning, R. C. (1966). *Lectures on Riemann Surfaces* (Princeton University Press, Princeton, New Jersey).
14. Brieskorn, E. (1964). *Math. Ann.*, **155**, 184.
15. Geroch, R. P. (1971). *Ann. Phys.*, **62**, 582.

[20] A modified form of Geroch's proposal [15] seems worth exploring here. I am grateful to A. Ashtekar for suggesting this.

71
THE NONLINEAR GRAVITON

Essay, awarded first prize, Gravity Research Foundation, 1975.
General Relativity and Gravitation **7**, no. 2 (1976), 171–6

This is the published version of the Prize Essay of Chapter 67. Curiously, its publication was delayed, so that it appeared after the publication of the later more complete article of Chapter 70.

26 The nonlinear graviton. Essay, awarded first prize, Gravity Research Foundation, 1975.
General Relativity and Gravitation **7**, no. 2 (1976), 171–6

The Nonlinear Graviton

ROGER PENROSE
Mathematical Institute, Oxford, England

Abstract

A new approach to quantized gravitational theory is suggested. It is argued by analogy with Maxwell theory—and also from a principle that (physical) gravitons should carry space-time curvature—that a free graviton should be describable by a complex solution of Einstein's vacuum equations. For a left-handed graviton (pure helicity state $-2\hbar$) we require a solution which is of a particular kind, called right-flat, and which is of positive frequency. A construction is given for obtaining all such solutions, in terms of general curved (top halves of) twistor spaces.

Despite impressive progress, recently [*1*], towards the intended goal of a satisfactory quantum theory of gravity, there remain fundamental problems whose solutions do not appear to be yet in sight. In particular, the standard perturbation techniques seem to lead to a quantized Einstein theory which is not renormalizable [*1*], and it has consequently been argued that Einstein's equations should perhaps be replaced by something more compatible with conventional quantum field theory. There is also the alternative possibility, which has occasionally been aired, that some of the basic principles of quantum mechanics may need to be called into question [*2*] if we are ever to arrive at a completely satisfactory quantum theory of gravity. It is this second alternative that I wish to explore here; and it is the principle of linear superposition (at least in its standard form) that I shall be calling into question.

Let us first ask: What is the most primitive physical requirement of a quantum theory? As the name implies, what we expect is that it should provide us with *quanta*—the particles associated with whatever field or fields we are attempting to describe. Thus, a quantum theory of gravity should provide *gravitons*. Furthermore, since we are trying to do physics, rather than mathematical

exercises, we require that a really physical concept of graviton actually does exist. For the purpose of this essay I am assuming this to be the case. I hope that the somewhat remarkable mathematical result that I shall give lends some independent theoretical support for this supposition.

I shall proceed by analogy with the electromagnetic field, at first. How, indeed, do we provide a mathematical description of a single free photon? I require a description which is both relativistically invariant and gauge invariant. The wave function can then be given as a solution F_{ab} $(= -F_{ba})$ of the free Maxwell equations

$$\nabla_{[a}F_{bc]} = 0, \qquad \nabla^{a}F_{ab} = 0 \tag{1}$$

However, there are two essential differences between classical solutions of the equations (1) and solutions giving wave functions for free photons. In the first place, classical solutions are *real* while wave functions are *complex*; in the second (and related to the first) place, wave functions should be composed only of *positive frequencies*.[1] There is also the question of choosing a normalization for the wave function, but I shall not consider that here, as it is not strictly necessary at this stage. But one can go further than this. The description so far is (in our world, where parity conservation is not a general law of nature) a reducible one. We may regard it as describing two *different* particles simultaneously, namely the left-handed and the right-handed photon. To examine this, consider the decomposition of F_{ab} into its self-dual and anti-self-dual parts:

$$F_{ab} = F_{ab}^{(+)} + F_{ab}^{(-)} \tag{2}$$

where

$$iF_{ab}^{(\pm)} = \pm\tfrac{1}{2}e_{abcd}F^{cd(\pm)}$$

e_{abcd} being the alternating tensor. If F_{ab} is real, $F_{ab}^{(+)}$ and $F_{ab}^{(-)}$ are complex conjugates of one another, but for a wave function, they are simply independent complex quantities, each of positive frequency (and therefore necessarily *not* complex conjugates of one another). In fact $F_{ab}^{(-)}$ describes a left-handed photon (helicity $-\hbar$) and $F_{ab}^{(+)}$ a right-handed photon (helicity $+\hbar$).

A spinor description can also be used. Put

$$F_{ab}^{(-)} = \phi_{AB}\epsilon_{A'B'} \qquad F_{ab}^{(+)} = \epsilon_{AB}\tilde{\phi}_{A'B'}$$

Then for a classical (real) field we have $\tilde{\phi}_{A'B'} = \overline{\phi}_{A'B'}$, whereas for a photon wave function ϕ_{AB} and $\tilde{\phi}_{A'B'}$ are independent (symmetric) spinors describing the left-

[1] There are, of course, many alternative equivalent ways of describing a free photon. The choice I am making here seems to me to be the most natural—and also the only one which fits in with the further ideas I am suggesting.

The nonlinear graviton. Essay, awarded first prize, Gravity Research Foundation, 1975. *General Relativity and Gravitation* **7**, no. 2 (1976), 171–6

handed and right-handed components of the photon [4]. The Maxwell equations (1) become

$$\nabla^{AA'}\phi_{AB} = 0 \qquad \nabla^{AA'}\tilde{\phi}_{A'B'} = 0 \tag{3}$$

To describe a pure $-\hbar$ helicity state we must take $\tilde{\phi}_{A'B'} = 0$; for a pure $+\hbar$ helicity state we take $\phi_{AB} = 0$. The requirement that ϕ_{AB} and $\tilde{\phi}_{A'B'}$ have positive frequency may be stated in the following geometrical way: It is possible to extend the domain of definition of each field quantity holomorphically to points of complexified Minkowski space $\mathbb{CM}$, whose position vectors $x^a + iy^a$ have imaginary part y^a lying on or within the past null cone [3]. This portion of $\mathbb{CM}$ will be denoted as $\mathbb{CM}^+$.

Let us now turn to the case of gravity. There is an obvious analogue of the above description, where two four-index symmetric spinors ϕ_{ABCD}, $\tilde{\phi}_{A'B'C'D'}$ are employed in a Minkowski background space to provide a (complex) linearized curvature tensor

$$K_{abcd} = \phi_{ABCD}\epsilon_{A'B'}\epsilon_{C'D'} + \epsilon_{AB}\epsilon_{CD}\tilde{\phi}_{A'B'C'D'}$$

The discussion is exactly as above. A common view of a graviton has, indeed, been essentially this, namely something which refers to weak-field gravitational perturbations of a space-time (normally flat). However, it is at this point that I wish to take issue with the standard view. I cannot (now) believe that a *physical* graviton should be described by linear gravitational theory. My reasoning is as follows: If we start with flat space, and then add one (*linearized*) "graviton," the space remains flat. The null cones have not shifted. We add a second such "graviton" and a third and a fourth, and the space is still flat with null cones still locked in their original Minkowskian positions. With such a perturbative viewpoint it is only after an infinite number of "gravitons" have been added that the space can become curved. The situation may be compared with a power series expansion. For any finite number of terms the function

$$1 + \frac{1}{z} + \frac{1}{z^2} + \cdots + \frac{1}{z^n}$$

has a pole singularity locked at the position $z = 0$. But the sum to infinity

$$1 + \frac{1}{z} + \frac{1}{z^2} + \frac{1}{z^3} + \cdots = \frac{z}{z-1}$$

has its pole shifted to the position $z = 1$. The usual perturbative expansions of graviton propagators have a similar behavior, with poles corresponding to the Minkowskian null cones right up until the sum to infinity has been performed. Surely, if it has meaning to talk about *physical* gravitons, it should have meaning to talk about finite numbers of such gravitons. And if Einstein's theory is physically appropriate, then surely *each graviton itself carries its measure of curvature*.

I cannot believe that curvature emerges only after an infinite sum has been performed.

So rather than working with linear theory, let us attempt to apply the electromagnetic analogy instead to the full nonlinear Einstein theory. Indeed, a decomposition of the Riemann curvature of a vacuum space-time can be expressed in a form analogous to (2)

$$R_{abcd} = \Psi_{ABCD}\epsilon_{A'B'}\epsilon_{C'D'} + \epsilon_{AB}\epsilon_{CD}\tilde{\Psi}_{A'B'C'D'}$$

where the two terms on the right are the anti-self-dual and self-dual parts of the vacuum curvature, respectively. For a classical (real) space-time, $\tilde{\Psi}_{A'B'C'D'}$ must be the complex conjugate of Ψ_{ABCD}. But to describe a (nonlinear) graviton we must expect, by analogy with the electromagnetic case, that Ψ_{ABCD} and $\tilde{\Psi}_{A'B'C'D'}$ will become independent quantities, each, in some appropriate sense, of positive frequency. The curvature R_{abcd} becomes complex, so we are dealing with some form of complex Riemannian space. The metric g_{ab} must, indeed, be complex symmetric (holomorphic, not Kähler). The complex Ricci tensor must vanish. The complex vacuum Bianchi identity becomes

$$\nabla^{AA'}\Psi_{ABCD} = 0 \qquad \nabla^{AA'}\tilde{\Psi}_{A'B'C'D'} = 0$$

in close analogy with (3).

But if such a situation describes a graviton it would generally be a graviton of mixed helicities. For a pure graviton in a helicity state we would apparently require $\tilde{\Psi}_{A'B'C'D'} = 0$ (helicity $-2\hbar$) or else $\Psi_{ABCD} = 0$ (helicity $+2\hbar$). I shall refer to a complex vacuum space-time for which $\tilde{\Psi}_{A'B'C'D'} = 0$ as *right-flat*; and one for which $\Psi_{ABCD} = 0$ as *left-flat*. Thus, I am proposing that a single left-handed graviton is described by a positive-frequency right-flat (left-curved) complex space-time; and a right-handed graviton by a left-flat space-time.[2]

Such a description may seem somewhat mysterious (and "positive frequency" has yet to be clearly defined in the context). But it is, to my mind, a very remarkable fact that the general such complex space-time can be constructed *automatically* by way of *twistor theory* [*4*, *1*]! There are already various reasons for believing that twistor theory may be able to provide some new insights into the workings of nature; and here we have an apparently independent role for this theory. It will not be possible to go into any detail, but the essential mathematical result can be briefly described.

Let us first recall [*4*] how complex (compactified) Minkowski space $\mathbb{CM}$ arises from standard (flat) twistor space $\mathbb{T}$. The space $\mathbb{T}$ is a complex four-dimensional vector space endowed with a Hermitian form Σ of signature $(++--)$. The parts of $\mathbb{T}$ at which Σ is positive, negative, or zero are labelled $\mathbb{T}^+$, $\mathbb{T}^-$, $\mathbb{N}$,

[2] The concepts of right-flat and left-flat space-times arose historically from E. T. Newman's construction of the spaces [*5*] referred to as 'Heavens'. Though much of this essay has its origins in Newman's construction, I shall not have need to refer to it explicitly here.

The nonlinear graviton. Essay, awarded first prize, Gravity Research Foundation, 1975. *General Relativity and Gravitation* **7**, no. 2 (1976), 171–6

respectively. The complex two-dimensional linear subspaces of $\mathbb{T}$ correspond to the points of $\mathbb{CM}$; those lying in $\mathbb{N}$ define the points of real (compactified) Minkowski space $\mathbb{M}$ those lying in $\mathbb{T}^+ \cup \mathbb{N}$ define the points of $\mathbb{CM}^+$. Often it is convenient to consider the projective twistor space $\mathbb{PT}$ (where the prefix $\mathbb{P}$ means 'projective'). The points of $\mathbb{CM}$ are then represented in $\mathbb{PT}$ by complex projective lines (compact holomorphic curves having topology S^2); the points of $\mathbb{M}$ by projective lines in $\mathbb{PN}$ and those of $\mathbb{CM}^+$ by projective lines in $\mathbb{PT}^+ \cup \mathbb{PN}$. The space $\mathbb{T}$ is also endowed with a (Poincaré invariant) 1-form and 2-form

$$\underline{\iota} = I_{\alpha\beta} Z^\alpha dZ^\beta \qquad 2\underline{\tau} = d\underline{\iota} = I_{\alpha\beta} dZ^\alpha \wedge dZ^\beta$$

(the form τ being degenerate, of rank 2); and with (a conformally invariant) 3-form and 4-form

$$6\underline{\rho} = \epsilon_{\alpha\beta\gamma\delta} Z^\alpha dZ^\beta \wedge dZ^\gamma \wedge dZ^\delta$$

$$24\omega = 6d\underline{\rho} = \epsilon_{\alpha\beta\gamma\delta} dZ^\alpha \wedge dZ^\beta \wedge dZ^\gamma \wedge dZ^\delta$$

in standard twistor notation [*1*].

To construct a left-handed nonlinear graviton we proceed as follows. We consider the bounded portion $\mathbb{PT}^+ \cup \mathbb{PN}$ of projective twistor space, deleting the part $\mathbb{PT}^-$. We then perturb its complex structure in an arbitrary fashion to obtain a new space $\mathcal{P}$. By appealing to a theorem of Kodaira [*6*] (and employing some observations due to M. F. Atiyah) we can infer the existence of a complex four-parameter family of compact holomorphic curves in $\mathcal{P}$, which are the analogues of the projective lines in $\mathbb{PT}$. Each such curve now defines a point in an abstract complex four-manifold $\mathfrak{M}$. The space $\mathfrak{M}$ naturally acquires a (complex) conformal structure, where null separation in $\mathfrak{M}$ corresponds to intersection of the corresponding curves in $\mathcal{P}$. This conformal structure can be shown to be Riemannian (by general theory) with a right-flat conformal curvature. We now require that the forms $\underline{\iota}$, $\underline{\tau}$, $\underline{\rho}$, $\underline{\omega}$ have been introduced, with the same stated properties as for flat twistors, but otherwise arbitrary. We apparently need solve no equations to achieve all this. But, remarkably, the resulting complex four-space $\mathfrak{M}$ acquires the structure of a general right-flat solution of the Einstein vacuum equations!

To ensure that $\mathfrak{M}$ is, in some appropriate sense, of positive frequency, we require that $\mathcal{P}$ is sufficiently extended that it is validly analogous to $\mathbb{PT}^+ \cup \mathbb{PN}$. I cannot explain the details here, but the requirement seems to be that the boundary of $\mathcal{P}$ is ruled appropriately by a real three-parameter family of holomorphic curves. Again, it seems that no equations need be solved to achieve this. Details will be published elsewhere.

There is clearly much more that needs to be done for this line of approach to quantum gravity to be made the basis of a physical theory. We require a definition of scalar product between these one-particle states. And how are we to

build up a many-graviton theory from this one-graviton approach?[3] Further work is in progress.

Acknowledgments

My thanks are especially due to E. T. Newman and M. F. Atiyah for supplying vital ideas.

Note Added in Proof

Owing to a delay incurred because this article had to be reset when this journal changed its format, an extended version containing some details has been published prior to this paper.

References

1. Isham, C. J., Penrose, R. and Sciama, D. W. (eds.) (1975). *Quantum Gravity: An Oxford Symposium*, (Oxford University Press).
2. Mielnik, B. (1974). *Commun. Math. Phys.*, **37**, 221.
3. Streater, R. F. and Wightman, A. S. (1964). *CPT, Spin and Statistics, and All That*, (W. A. (Benjamin, Inc., New York).
4. Penrose, R. and MacCallum, M. A. H. (1973). *Phys. Repts*, **C6**, 242.
5. Newman, E. T. (1975). *General Relativity and Gravitation*. John Wiley and Sons, New York.
6. Kodaira, K. (1963). *Amer. J. Math.*, **85**, 79.
7. Geroch, R. P. (1972). *Ann. Phys.*, **62**, 582.
8. Penrose, R. (1976). *Gen. Rel. Grav.*, **7**, 31.

[3] A modified form of Geroch's proposal [7] seems worth exploring here. I am grateful to A. Ashtekar for suggesting this.

72
THE SPACE-TIME SINGULARITIES OF COSMOLOGY AND IN BLACK HOLES

IAU Symposium Proceedings Series No. **13**, *Cosmology, 1976*

34 The space-time singularities of cosmology and in black holes. *IAU Symposium Proceedings Series No.* **13**, *Cosmology, 1976*

The Space-Time Singularities of Cosmology and in Black Holes

R. Penrose
(Mathematical Institute, Oxford)

The mathematical theory of cosmology is based, in the first instance, on certain solutions of Einstein's equations known as cosmological models [1]. These models assume a high degree of symmetry. This is necessary mathematically because Einstein's equations are complicated and nonlinear, so without such symmetry assumptions the equations tend to become intractable. But it also appears, fortunately, that such symmetry and uniformity is actually true of the observed universe to a very high degree of approximation (as is indicated, particularly, by the observed isotropy - to at least one part in a thousand - of the 2.7°K microwave background radiation).

A striking feature of these cosmological models is the presence of an initial singularity - representing the "big bang" origin of the universe. In some (but not all) models there is recollapse also to a final singularity. From the observational point of view, as we have heard in Dr. Sciama's paper, there is strong evidence (from the 2.7°K radiation

again) that the universe was once in a highly condensed state. The density and general curvature of space-time must have been a great many orders of magnitude larger about 10^{10} years ago than it is now. (There is no clear evidence, on the other hand, that the universe will ever recollapse to a final singularity.) If we extrapolate backwards in time and assume that the normal homogeneous and isotropic type of cosmological model remains a good approximation, then we are led to the picture of ever increasing densities and space-time curvatures, as an initial singularity is approached. At the singularity itself, the model would possess infinite densities and curvatures.

But have we any good reason to believe that the initial singularity of the cosmological model represents a true feature of the actual universe? Since high symmetry has been assumed for the model, it is perhaps not surprising that the extrapolation backwards in time leads to infinities. All the matter in the world is focused inwards to one tiny point, so the presence of an infinite density in the model might be attributed merely to the exactness of this focusing. We may envisage, instead, that small irregularities in the matter distribution (which indeed are present in the observed universe), when extrapolated backwards could destroy the exactness of the focusing. Instead, we might imagine a very large but finite density in an exceedingly complicated and asymmetrical condensed state - this state having possibly arisen from the asymmetrical collapse of a previous contracting phase of the universe.

The space-time singularities of cosmology and in black holes. *IAU Symposium Proceedings Series No.* **13**, *Cosmology, 1976*

Such a picture had indeed been suggested by a number of authors, up until about 10 years ago. But since that time, certain mathematical theorems, known as singularity theorems [1], [2], have been established which show, in effect, that no such asymmetrical "bounce" avoiding an actual singularity can ever be achieved strictly within classical general relativity theory. In somewhat more precise terms the strongest (effectively) of these theorems states [2]:

A space-time must contain a singularity if it

(*i*) satisfies Einstein's equations with energy condition,

(*ii*) contains no closed timelike curves,

(*iii*) is sufficiently general and

(*iv*) contains either

(*a*) a closed spacelike hypersurface,

(*b*) a trapped surface,

or (*c*) a point whose light cone begins to reconverge.

Some explanation of these conditions is required. In (*i*), Einstein's equations are taken without cosmological term

$$R_{ab} - \tfrac{1}{2}Rg_{ab} = -8\pi G T_{ab}$$

(although the result would still hold with negative cosmological constant - and there are other theorems which can cope also with a cosmological constant with either sign). The energy condition states a "physically reasonable" condition to be satisfied by the matter tensor T_{ab}. This asserts, in effect, that the local energy density is non-negative definite for all observers and that the sum of the principal

pressures and energy density (in relativistic units) is everywhere non-negative. This is a very reasonable requirement for classical (non-quantum) matter and is easily satisfied for all equations of state that are seriously considered by astrophysicists. Condition (*ii*) states that no material particle can travel in such a way that it can enter its own past. A violation of (*ii*) would lead to the science-fiction possibility that a space traveller could return to earth before he was born and so persuade his parents never to meet. Space-times with this property are normally considered to be physically unreasonable! Condition (*iii*) can be stated more precisely in mathematical terms to the effect that there is a certain amount of curvature somewhere along each timelike or null (= lightlike) geodesic. It represents no real physical restriction on the space-time model since some irregularities are always to be expected. But it is needed for the mathematical statement of the theorem.

All of (*i*), (*ii*) and (*iii*) must be considered to be very reasonable requirements for any realistic classical general-relativistic space-time model. It is condition (*iv*) (in one of its three versions) that states that the model describes a sufficiently extreme (global) situation that space-time singularities are inevitable. Such a situation occurs, according to (*iva*), simply if the universe is spatially closed. The condition (*ivb*) is a criterion which can be used to signify that a "point of no return" has taken place in a gravitational collapse of a star

The space-time singularities of cosmology and in black holes. *IAU Symposium Proceedings Series No.* **13**, *Cosmology, 1976*

(or collection of stars or of gas, etc.) leading to the probable formation of a black hole [1], [3], [4]. I shall not bother with the technical definition of (*ivb*) here. Condition (*ivo*) is somewhat similar in content to (*ivb*) but a little easier to state. It can be used either in the forward time direction (to characterise a collapse past a "point of no return") or in the backward time direction (to indicate the presence of an explosive origin to the universe). In the forward version, (*ivo*) states that some flash of (idealised) light emanating from some space-time point has the property that the expansion of the light rays slows down and turns into a contraction (in every direction). This would be expected to occur if sufficient collapsing matter falls inwards across the light rays towards the centre. This focusing effect of matter on light rays is, in fact, one of the <u>observed</u> phenomena of general relativity - the bending of light by the sun. What is required is simply that sufficient such matter collapses to a sufficiently small region that the focusing changes the expansion of the rays to a contraction. Such a situation is expected to occur in the final collapse of a very massive star after its internal resources of energy are exhausted, a black hole being expected as the final result. The backward version of (*ivo*), on the other hand, concerns an inward flash of light converging on a space-time point. If this point is considered to be the earth at the present time, then the "light flash" (*i.e.* the past light cone) consists of the history of all observed photons entering the earth from the outside world. The statement

(*ivc*) that this light cone reconverges, taken in the past sense, has the following interpretation. If an object of fixed diameter is removed to greater and greater distances from the earth, it presents, to begin with, a smaller and smaller apparent size (angular diameter). But this is true only down to a certain minimum apparent size if the reconverging light cone effect takes place. As the object is removed yet farther away (and consequently also farther into the past in order that it remain visible to us now) its apparent size begins to increase again. This cosmological effect is directly observable in principle - and even perhaps in practice [5]. There are good theoretical reasons for believing that the effect must actually be true of our universe.

The conclusion of the theorem - that the space-time is singular - also requires some elaboration. The singularity theorems do not, unfortunately, say anything about space-time curvatures (or densities) approaching infinity. They only say that, in an appropriate sense, the space-time has to be incomplete. This means that certain freely moving material particles or photons have world lines which cannot be extended indefinitely in the space-time. In effect, the space-time has to "come to an end" somewhere either in the past or the future. This "coming to an end" is how we are to view the presence of a singularity. In realistic models, this presumably arises because space-time curvatures diverge to infinity, but this actual divergence of curvature has not yet been mathematically established for the general case.

Assuming that divergence of curvature is indeed the true reason for the space-time's incompleteness, we may ask what this means physically. There are two situations to consider, namely the big bang initial singularity of the universe and the singularities arising in gravitational collapse. Roughly speaking, these situations are time-reverses of one another. However the collapses which are envisaged need not comprise the entire universe, but may be the more local collapses of stars (or star systems, etc.) to form black holes [1], [3], [4]. There are good theoretical and observational reasons for believing that black holes should actually occur fairly frequently in the universe. Large stars are observed of, say, 20 to 100 times the mass of the sun, and it seems that the normal evolution of such a star should result in a final collapse to a black hole. The actual matter of the collapsing star apparently gets crushed to ever increasing densities, its ultimate fate being annihilation at the space-time singularity at the black hole's centre. This reflects the time-reversed behaviour of the initial big bang singularity, which represents the initial creation of matter in the universe.

The picture of a space-time singularity that has been presented here is as the initial or final state of the universe as we know it. The singularities are consequences of physical theory as it is presently understood. They represent regions of breakdown of existing theory - their occurrencesbeing actually consequences of this same theory. This is not to say that physics will never be able to explain

what happens at such regions. Indeed, it is strongly to be hoped that some future theory will be able to handle them. Perhaps some general scheme incorporating both general relativity and quantum mechanics will be forthcoming in the not-too-distant future, which hopefully will have something useful to say about the space-time singularities. This would be important particularly for the big bang, since the observed features of our universe ultimately depend upon the detailed nature of this big bang. On the other hand, the structure of the singularity inside a black hole would not, in the normal view of things, have any observable effects.

There is an alternative possibility which is sometimes envisaged, however, concerning the singularities arising in gravitational collapse. It might be that a black hole is not always the result but, instead, the final singularity might sometimes remain visible to the outside world. Such a singularity would be called a naked singularity, but it is not yet known whether they are permitted by existing theory. If they are, then we should have a still stronger reason for requiring a detailed new theory for handling singularities. Unlike the space-time singularities in black holes, naked singularities would (if they exist) have definite observable effects. It should, however, be mentioned, for what it is worth, that the present concensus among theoreticians seems to be rather against the possibility that naked singularities can occur.

Setting aside the problem of the existence of naked singularities, there is a further intriguing question that

The space-time singularities of cosmology and in black holes. *IAU Symposium Proceedings Series No.* **13**, *Cosmology, 1976*

can be asked. It may be recalled that all the laws of local physics - apart from one very minor exception, namely K^0-meson decay [6], which seems totally irrelevant to all physical processes of importance - are symmetrical in time. That is, any motion picture film depicting an allowable physical process (involving, say, particle collisions, electromagnetic or gravitational interactions, or strong or weak nuclear forces) could be run equally well backwards or forwards and the physical laws would still hold good. However this does not remain valid, in effect, when we take large scale statistical phenomena into account. Thus, some processes are found to occur much more frequently in one direction in time than in the other. This asymmetry is associated with the statistical quantity called the entropy of a system, which is found to increase in the positive direction of time. The normal attitude to this is that the entropy increase phenomenon is determined by the boundary conditions of the universe, not the local physical laws. Thus, the universe is viewed as starting off, for some reason, in a state of comparatively low entropy, and it ends up in a state of high entropy. But what is the initial state of the universe? - apparently the big bang singularity. And the final state? For matter inside a black hole it is still a singularity - although it may well be that for matter outside an indefinite expansion is more probable. But in any case we must expect a difference in entropy between the state at an initial singularity and that at a final singularity, the entropy being low in the first case and high in the second. Does this difference show up in the geometric structure of the singularity? I think it is very likely that

it does.

To picture what this difference might be, we envisage an observer falling into a black hole. He apparently encounters space-time curvatures mounting to infinity, which he feels as an inexorably mounting tidal force. This tidal force stretches him in one direction and squashes him in others, but in essence it is a volume preserving effect. This tidal force arises from what is called the Weyl curvature [1] of space-time. On the other hand, the large space-time curvature near the big bang appears to be of quite a different kind. An observer would feel himself to be squashed uniformly in all directions at once. This kind of curvature force arises from the Ricci curvature of space-time. In general, the total space-time curvature (Riemann curvature) is composed of both parts (the Ricci part, according to Einstein's equations, being directly related to the matter density). But it seems that near the initial singularity it is Ricci curvature which dominates, while at a final singularity Weyl curvature dominates - suggesting that some measure of entropy in space-time structure may be related to the Weyl curvature.

Here, it seems, there is a gross time-asymmetry in local physical behaviour. It may well be that whatever new physics is needed for handling space-time singularities will need to be asymmetrical in time. Perhaps we are witnessing a small residual effect of this in the time-asymmetric behaviour of the K^0-particle decays. Only time - and a new theory - will tell!

REFERENCES

1. Hawking, S.W., and Ellis, G.F.R., "The Large Scale Structure of Space-Time," CUP, 1973.
2. Hawking, S.W., and Penrose, R., *Proc. Roy. Soc. A*, 1970 314, 529.
3. Misner, C.W., Thorne, K.S., and Wheeler, J.A., "Gravitation," W.H. Freeman & Co., San Francisco, 1973.
4. Penrose, R., "Black Holes," *in* John, L., *Editor* "Cosmology Now," BBC publication, London 1973.
5. Reedhead, A.C.S, and Hewish, A., *Mem. Roy. Astr. Soc.*, 1974, 78, 1.
6. Casella, R.S., *Phys. Rev. Letters*, 1968, 21, 1128.

73
THE TWISTOR QUADRILLE

Twistor Newsletter **1** (1976). With G. A. J. Sparling. Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

This procedure of this article refers to the Ward construction of anti-self-dual Maxwell fields, rather than to the non-linear graviton construction of Chapters 67, 70, and 71 for anti-self-dual Einstein fields. In the electromagnetic case, the Čech-type procedure of piecing together complex manifolds out of patches of twistor space still applies, and is here extended to the seemingly simple case of an anti-self-dual electromagnetic field with a source. It turns out that there are lots of patches! The integrality of charge that occurs here can also be understood directly in terms of exact sequences, etc. See pp. [34] and[35] of Chapter 77, and also Chapter 109. There is, however, an important subtlety that we had not noticed at this stage. This is explained in Chapter 108 and generalized in Chapters 134 and 135.

46 The Twistor Quadrille. *Twistor Newsletter* **1** (1976). With G. A. J. Sparling. Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

§3.3. THE TWISTOR QUADRILLE: A LINE BUNDLE BASED ON THE COULOMB FIELD

by R. Penrose and G.A.J. Sparling

The aim of this article is to show explicitly how to construct a twistor space which encapsulates the global structure of a left-handed (i.e. anti-self-dual) Coulomb field. The construction dovetails in neatly with an analogous construction for the Schwarzschild-like graviton (see §3.4). We carry out the construction of §3.2 for a line bundle over part of projective twistor space incorporating the Coulomb field in any suitable sufficiently small four-dimensional neighbourhood (with topology C^4), of complexified Minkowski space, on which the field is non-singular. We then try to extend this line bundle to handle the Coulomb field in a complex four-dimensional neighbourhood of a two-dimensional real sphere, surrounding the source singularity. We find that our construction breaks down unless a certain integrality condition holds; the construction forces electric charge to come in fixed multiples of one basic unit. This is closely related to the famous Dirac argument (which uses magnetic monopoles) but is not quite the same.

The construction goes as follows. We first recall the building of the line bundle over that portion of projective twistor space which contains all the projective twistors which pass through points belonging to a neighbourhood with topology C^4 in complex Minkowski space. This portion may be regarded as a certain neighbourhood U, say, of a line in projective twistor space. As usual we split the neighbourhood into two overlapping regions: U_1, with projective coordinates $\{Z_1^\alpha\}$, and U_2, with projective coordinates $\{Z_2^\alpha\}$, such that the overlap $U_1 \cap U_2$ contains an annular region of each compact projective line in U. We also use coordinates Z_i^α for that part of $C^4-\{0\}$ which projects down to U_i under the canonical map $C^4-\{0\} \to CP^3$, i=1,2.

We may assume that the projective coordinates $\{Z_1^\alpha\}$ for a point on $U_1 \cap U_2$ are the same as $\{Z_2^\alpha\}$. Then if we identified Z_1^α with Z_2^α on the overlap region we would obtain just the usual non-projective twistors which project to U, as a principal bundle of a line bundle over U. However here we identify the point with coordinates Z^α in the U_2 chart with the point with coordinates $\exp\{f(Z^\beta)\}\, Z^\alpha$ in the U_1 chart (i.e. $Z_1^\alpha = Z^\alpha$ corresponds to $Z_2^\alpha = \exp\{f(Z^\beta)\}Z^\alpha$), where we take

$$f(Z^\alpha) = \log\left(\frac{Q}{Z^2\, Z^3}\right) ,$$

$$Q = Z^1\, Z^2 - Z^0\, Z^3 .$$

So our overlap region has to be such that it avoids the singularities $Q = 0$, $Z^2 = 0$ or $Z^3 = 0$, appropriately. This function f produces a Coulomb field when it is substituted into the familiar contour integral formula.

To study the overlap region more carefully, we picture the singularities of the twistor function in Minkowski space. We may restrict our attention to a real two-sphere surrounding the charge, as the picture will be essentially unchanged in a sufficiently small complex neighbourhood of that sphere. We may assume that the sphere lies in a t=constant plane with centre on the charge wordline in a frame in which the charge is at rest. At each point $x^{AA'}$ of the two-sphere there are (generically) four spinors $\pi_{A'}$, defined up to proportionality, such that if $Z^\alpha = (i\, x^{AA'}\, \pi_A, \pi_{A'})$, then $f(Z^\alpha)$ is singular. We project the null vectors $\bar{\pi}^A\, \pi^{A'}$ into the surface t=constant and obtain the picture in Fig. A, the flag-ended lines representing the singular directions at each point. The flags A and B at each point correspond to the singularities $Z^2 = 0$ and $Z^3 = 0$ respectively. The flags O and I correspond to the singularities $Q = 0$, O pointing outward from and I pointing inward to the

 The Twistor Quadrille. *Twistor Newsletter* **1** (1976). With G. A. J. Sparling. Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

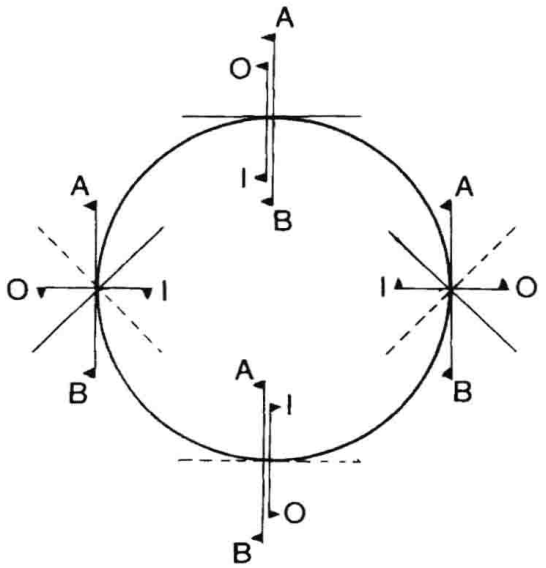


Figure A

centre. For simplicity and with no real loss of generality we may assume that A coincides with O at the north pole, and B with I. Then the whole picture is rotationally symmetric about the north-south axis. We portray the overlap region at various points of the sphere by drawing a solid line to represent the orientation, in the sphere of directions at each point, of the central great circle of the overlap region, for the sphere's worth of twistors that represent the given point in twistor space. (In this construction the overlap region may always be taken to consist of an annular neighbourhood of a great circle of the sphere of directions, for each point.) At the north pole the annular region may be taken to be equatorial as indicated in Fig. A. As we move away from the north pole, O moves away from

A and the annular region gradually has to tip up to allow for this. Eventually when we reach the south pole, the region becomes pinched. We may define the region U_1 by saying that for each point on the partial sphere in which a neighbourhood of the south pole is deleted, we have a partial sphere of directions for the twistors through those points, this partial sphere being tipped up at the point x on the equator (given the tipping up at x, we may work it out everywhere else by continuity). This is illustrated in Fig. B, where the sets U_i and their intersections are represented in the

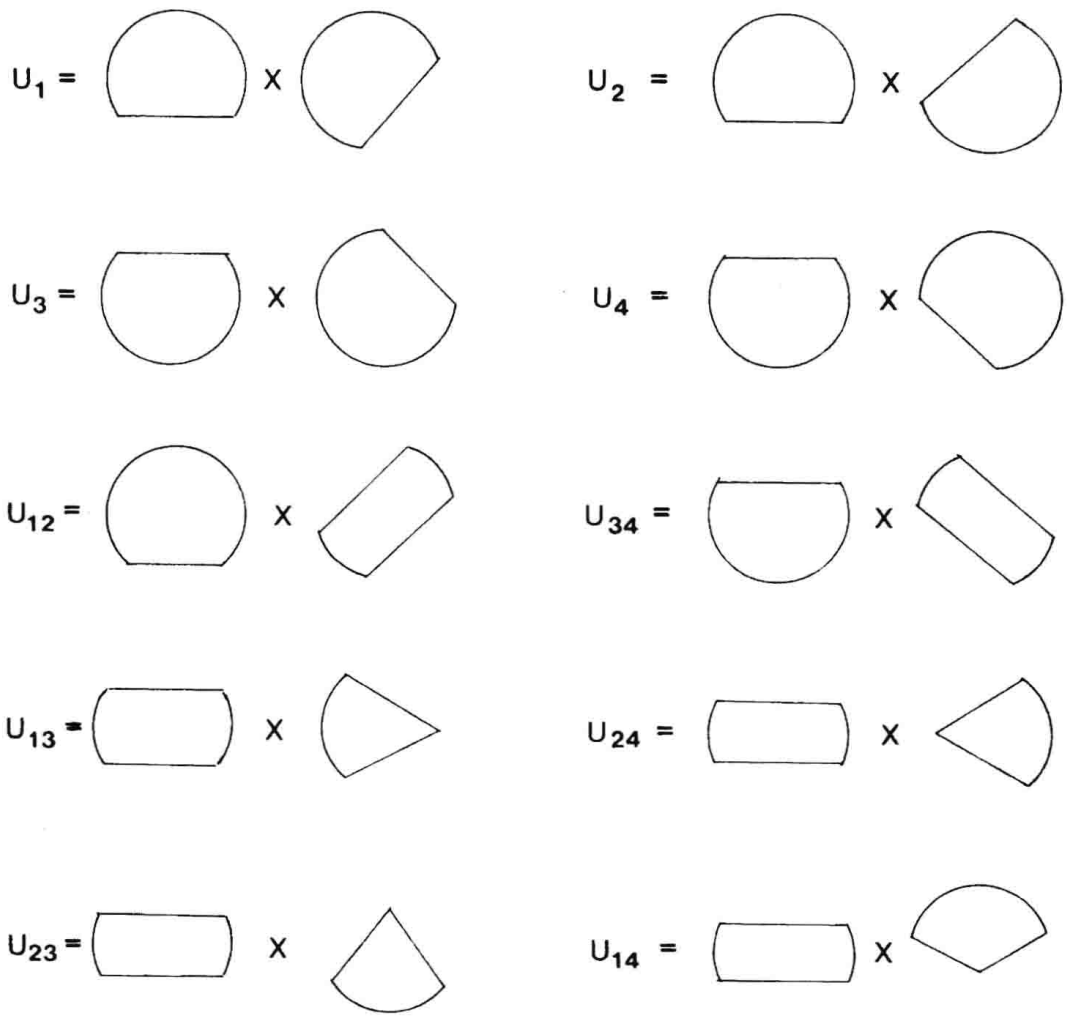


Figure B

The Twistor Quadrille. *Twistor Newsletter* **1** (1976). With G. A. J. Sparling. Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

form U_i = (partial sphere of points) × (allowed directions at a point x on the equator). The transition function to be used on the overlap region is $c_{12}(Z^\alpha) = \exp\{f(Z^\alpha)\} = Q\,(Z^2\,Z^3)^{-1}$.

This finishes the construction as far as the northern hemisphere is concerned and produces a twistor space carrying the structure of the Coulomb field in a small neighbourhood of that hemisphere. However we could equally well start from the south pole, using regions U_3 and U_4 as indicated with a dotted line in Fig. A and illustrated in Fig. B. The coordinates Z_3^α in U_3 and Z_4^α in U_4 are patched together according to $Z_3^\alpha = c_{34}(Z^\beta)\,Z_4^\alpha$ on the overlap region $U_{34} = U_3 \cap U_4$, where $c_{34}(Z^\alpha) = 0\,(Z^2\,Z^3)^{-1}$. Now can we patch all the patches together to give a line bundle corresponding to a neighbourhood of the whole two-sphere? We *can* do so if we can find transition functions $c_{ij}(Z^\alpha)$ defined on $U_{ij} = U_i \cap U_j$ such that $c_{ij}\,c_{jk} = c_{ik}$ on $U_{ijk} = U_i \cap U_j \cap U_k$, for all i,j,k (with $c_{ii} = 1$, $c_{ij} = c_{ji}^{-1}$), identifying $Z_j^\alpha = Z^\alpha$ in U_j with $Z_i^\alpha = c_{ij}(Z^\beta)\,Z^\alpha$ in U_i.

Now consider, for instance, the relation $c_{12} = c_{14}\,c_{42}$ on U_{124}, so $Q(Z^2\,Z^3)^{-1} = c_{14}\,c_{42}$. Now U_{14} contains the direction $Z^2 = 0$, so c_{14} cannot be singular there. However U_{14} does not contain directions where $Q = 0$ or $Z^3 = 0$. Also U_{24} contains a direction where $0 = 0$ but not where $Z^2 = 0$ or $Z^3 = 0$. So we may solve for

$$c_{14}(Z) = Q\,(Z^3)^{-2} \quad \text{and} \quad c_{42} = Z^3\,(Z^2)^{-1}.$$

Next consider $c_{12} = c_{13}\,c_{32}$ on U_{123}. A similar analysis yields

$$c_{32} = Q\,(Z^2)^{-2} \quad \text{and} \quad c_{13} = Z^2\,(Z^3)^{-1}.$$

Then we have $c_{32}\,c_{34} = c_{34}$ on U_{234} and $c_{31}\,c_{14} = c_{34}$ on U_{134}, so we have a complete solution. The four singularities are just right that they can

dance out of trouble. The solution is unique up to multiplication by an overall constant. To see the integrality of charge showing up, we look at the effect of replacing $f(Z)$ by $\lambda f(Z)$ for some $\lambda \neq 0$, $\lambda \in \mathbb{C}$. Then $c_{ij}(Z)$ becomes $(c_{ij}(Z))^{\lambda}$ and again this is the only possible solution up to an overall constant. But then $(c_{13})^{\lambda}$ is not single-valued on U_{13} nor is $(c_{24})^{\lambda}$ on U_{24}, unless λ is integral. So the construction fails to work globally around the two-sphere unless λ is integral. Although we have not proved here that no other construction will work, this follows from a sheaf-cohomological discussion (see §2.2).

74
THE SNAIL CONTOUR

Twistor Newsletter **1** (1976)

The fact that this article appeared in the same *Twistor Newsletter* (𝕋ℕ1) as Chapter 73 indicates that although we had understood the Čech-type, potentially non-linear, piecing together of twistor manfolds, we had—curiously enough, at that time—not yet understood the Čech-type *linear* sheaf-cohomological underpinnings of the twistor contour integral. Though the subsequent cohomological viewpoint (Chapter 77) renders this particular article an 'historical curiosity', it is nevertheless a rather curious curiosity, possibly not without some significance.

10. The Snail Contour

(working with Roger Penrose & Brian Branson)

This is in aid of an understanding of the inverse twistor function integral (due to B.D.B., G.A.J.S. & R.P.) described in outline on pp. 311–316 of "Quantum Gravity". But the snail contour would seem to have relevance in other contexts too.

The basic problem is as follows: we have an analytic function F defined on the real sphere S^2 and we wish to re-express the integral $\oint_{S^2} F\,dS$ (dS = 2-surface area element) as a contour integral (1-dim., over S^1) of a holomorphic function f which is obtained by a 1-dimensional complex definite integral of F. More explicitly

$$I = \oint_{S^2} F\,dS = \oint_{\zeta\in S^2} F(\zeta,\bar\zeta)\,\frac{4\,d\zeta\wedge d\bar\zeta}{(1+\zeta\bar\zeta)^2} = \oint_{\mathbb{S}^2} G(\zeta,\bar\zeta)\,d\zeta\wedge d\bar\zeta = \oint_{\Sigma} G(\zeta,\eta)\,d\zeta\wedge d\eta$$

over contour homologous to $\eta=\bar\zeta$

and the question is, can we express

$$I = \oint_\gamma d\zeta \int_{p(\zeta)}^{q(\zeta)} G(\zeta,\eta)\,d\eta = \oint_\gamma f(\zeta)\,d\zeta\,,$$

γ is some suitable S^1 contour on the ζ-sphere

with $f(\zeta) = \int_{p(\zeta)}^{q(\zeta)} G(\zeta,\eta)\,d\eta$?

$G(\zeta,\bar\zeta) = \dfrac{4\,F(\zeta,\bar\zeta)}{(1+\zeta\bar\zeta)^2}$ with G, F real-analytic; $G(\zeta,\eta)$, $F(\zeta,\eta)$ holomorphic.

Now if $F=1$ it's easy. We can take $p=0$ and $q=\infty$, and all is simple. This works because the singularities of G on the complexified sphere $\mathbb{C}S^2 \cong S^2\times S^2$ are "a long way away" from the real slice S^2. To represent a point of $\mathbb{C}S^2$, we take two independent points $\zeta, \bar\eta \in S^2$. The real points of S^2 are given when $\bar\eta=\zeta$. The particular function $G(\zeta,\eta) = 4(1+\zeta\eta)^{-2}$ (corr. $F=1$) has its singularities given when $\zeta=-1/\eta$, i.e. when ζ and $\bar\eta$ are antipodal points. For "rougher" functions G, the singularities occur when ζ and $\bar\eta$ are closer together on S^2. For a given analytic G there will be an $\epsilon>0$ such that $G(\zeta,\eta)$ is regular whenever ζ and $\bar\eta$ are closer than ϵ on the sphere S^2. We can think of ϵ as measuring a degree of "calmness" of G.

The idea is as follows. First fix γ as some closed real (S^1) contour on S^2. Define a 2-dimensional contour Σ, homologous to the real S^2 in $\mathbb{C}S^2 - \{$singularity set of $G\}$, thus:

$\Sigma = D_1 \cup C \cup D_2$ ⟶

(D_1 & D_2 are two disc-like regions on Σ and C is the remaining annular cylinder)

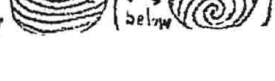
D_1 is defined as ζ moves about one side S_1 of γ ($\subset S^2$) and $\bar\eta = \overline{p(\zeta)}$ follows it about (anti-holomorphically)

D_2 " " " ζ " " the other " S_2 " γ " " " $\bar\eta = \overline{q(\zeta)}$ " " " "

C " " " ζ " along γ and $\bar\eta$ moves independently directly from $\overline{p(\zeta)}$ to $\overline{q(\zeta)}$ for each ζ.

We have $I = \oint_\Sigma G(\zeta,\eta)\,d\zeta\wedge d\eta = \int_{D_1} G\,d\zeta\wedge d\eta + \int_C G\,d\zeta\wedge d\eta + \int_{D_2} G\,d\zeta\wedge d\eta$. But $d\zeta\wedge d(p(\zeta)) = 0 = d\zeta\wedge d(q(\zeta))$.

$\therefore \int_{D_1} = 0 = \int_{D_2}$, so $I = \int_C G\,d\zeta\wedge d\eta = \oint_\gamma d\zeta \int_{p(\zeta)}^{q(\zeta)} G(\zeta,\eta)\,d\eta$, as required, provided that G is regular on D_1, D_2 and C, i.e. provided that $\overline{p(\zeta)}$ is closer to ζ than ϵ on D_1 and that $\overline{q(\zeta)}$ is closer to ζ than ϵ on D_2.

If ϵ is small, we need an elaborate-looking γ which winds closely all over S^2, such as ⟶ 

Then:

75
THE UNIVERSAL BRACKET FACTOR

Twistor Newsletter **1** (1976). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

One of the important contributions that the great Leonhard Euler has made to twistor theory! Such things feature strongly in Andrew Hodges's later work, now acquiring considerable importance in current understandings (as of 2010) of particle scattering theory.

The Universal Bracket Factor. *Twistor Newsletter* **1** (1976). Also published in *Advances in Twistor Theory.* Ed. L P Hughston, Pitman, 1979

§5.2 THE UNIVERSAL BRACKET FACTOR *by R. Penrose*

The standard bracket factor for twistor integrals is:

$$(x)_n = \begin{cases} \dfrac{-(n-1)!}{2\pi i(-x)^n} & \text{if } n = 1,2,3... \qquad (\oint) \\ \dfrac{x^{-n}}{(-n)!} & \text{if } n = 0,-1,-2,-3... \qquad (\oint\!\!\!\!\!\!-) \end{cases} .$$

The symbol $\oint$ means that the expression is to be inserted in an integral with a closed contour surrounding $x = 0$, whereas the symbol $\oint\!\!\!\!\!\!-$ means that the expression is to be integrated over an open contour with boundary in $x = 0$.

The *universal bracket factor* is

$$[x] = \sum_{n=-\infty}^{\infty} (x)_n ,$$

but this expression must be treated as *formal* at the moment (and a rigorous definition from someone would be welcome) because

$$2\pi i \sum_{n=1}^{\infty} (x)_n \overset{(\oint)}{=} \frac{1!}{x^2} + \frac{2!}{x^3} - \frac{3!}{x^4} + \ldots ,$$

which *diverges* for *all* x. However, as is described in Hardy 1973 pp.26-29, Euler was formally able to derive

$$\frac{1}{x} - \frac{1!}{x^2} + \frac{2!}{x^3} - \frac{3!}{x^4} + \ldots = e^x\{-\gamma - \log x + x - \frac{x^2}{2.2!} + \frac{x^3}{3.3!} - \ldots\},$$

which is valid with the left hand side an asymptotic series (γ = Euler's constant). Evaluating the right hand side we obtain

$$\frac{1}{x} - \frac{1!}{x^2} + \frac{2!}{x^3} - \frac{3!}{x^4} + \ldots = \int_0^\infty \frac{e^{-w}}{x+w}\, dw ,$$

which is a perfectly good function! But, unfortunately, $x = 0$ is now a branch

point and not a pole, so $(\oint)$ of this expression is not strictly meaningful as it stands. Note that

$$\sum_{n=-\infty}^{0} (x)_n \overset{(\oint)}{=} 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots = e^x = \frac{1}{2\pi i} \oint \frac{e^{-w}}{x+w}\, dw \ .$$

So (formally) we obtain

$$[x] = \{(\oint) \int_0^\infty + (\not\oint) \oint \} \frac{e^{-w}}{x+w}\, dw \ ,$$

an observation due to G.A.J. Sparling. We have (formally) $d[x]/dx = [x]$ (with suitable contour). Thus, *irrespective of homogeneities*, we have the following formula for the "twistor transform":

$$\tilde{f}(W_\alpha) = \oint [W_\alpha Z^\alpha] f(Z^\alpha)\, d^4Z;$$

and similarly for the scalar product we have:

$$\langle f|g\rangle = \frac{1}{(2\pi i)^6} \oint f(Z^\alpha)[Z^\alpha W_\alpha] g(W_\alpha)\, d^4Z \wedge d^4W \ .$$

The universal bracket $[Z^\alpha W_\alpha]$ has many (formal) uses, e.g. in twistor diagrams.

REFERENCE

Hardy, G.H. 1973 *Divergent Series*. Clarendon Press: Oxford.

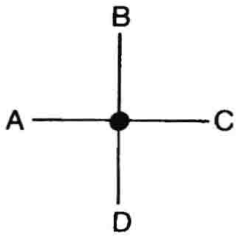
76
SOME ELEMENTARY TWISTOR INTEGRALS WITH BOUNDARY

Twistor Newsletter **1** (1976). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

Simple things, basic to twistor diagram theory, but also of current interest!

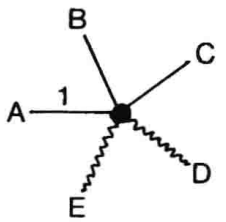
Some Elementary Twistor Integrals with Boundary. *Twistor Newsletter* **1** (1976). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

§5.4 SOME ELEMENTARY TWISTOR INTEGRALS WITH BOUNDARY *by R. Penrose*



$$\propto \oint \frac{DZ}{A\ B\ C\ D \atop Z\ Z\ Z\ Z} \propto \frac{1}{A\ B\ C\ D}$$

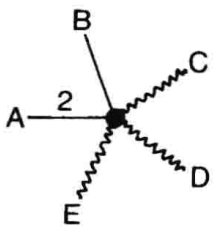
Figure A



$$\propto \oint \frac{DZ}{\left\{ {D \atop Z} \ {E \atop Z} \right\} \left({A \atop Z} \right)^2 \ {B \atop Z} \ {C \atop Z}} \propto \frac{B\ C\ D\ E}{A\ B\ C\ D \quad A\ B\ C\ E}$$

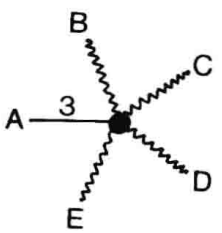
Figure B

$$\propto \oint \frac{\log \left({D \atop Z} / {E \atop Z} \right) DZ}{\left({A \atop Z} \right)^2 \ {B \atop Z} \ {C \atop Z}}$$



$$\propto \oint \frac{DZ}{\left\{ {C \atop Z} \ {D \atop Z} \ {E \atop Z} \right\} \left({A \atop Z} \right)^3 \ {B \atop Z}} \propto \frac{\left(B\ C\ D\ E \right)^2}{A\ B\ C\ D \quad A\ B\ C\ E \quad A\ B\ D\ E}$$

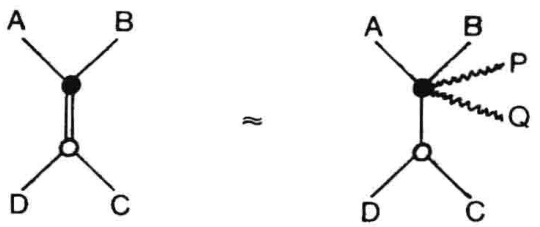
Figure C



$$\propto \oint \frac{DZ}{\left\{ {B \atop Z} \ {C \atop Z} \ {D \atop Z} \ {E \atop Z} \right\} \left({A \atop Z} \right)^4} \propto \frac{\left(B\ C\ D\ E \right)^3}{A\ B\ C\ D \quad A\ B\ C\ E \quad A\ B\ D\ E \quad A\ C\ D\ E}$$

Figure D

In Figure A one finds the standard twistor vertex, as described in Penrose and MacCallum 1972. In Figures B, C and D one finds example of the evalution of some twistor integrals with boundary. For details of the notation involved here see §5.1. Note that the "4 lines at a vertex" rule is violated. G.A.J. Sparling noticed some years ago that



where the second diagram is evaluated by doing the filled vertex first, and the unfilled vertex second. Similarly, we have:

REFERENCES

Penrose, R. & MacCallum, M.A.H. 1972 Phys. Reports 6, 241-315.

Penrose, R. 1975 *Twistor Theory, its Aims and Achievements.* In: *Quantum Gravity,* eds. C.J. Isham, R. Penrose, & D.W. Sciama, pp. 268-407. Clarendon Press: Oxford.

77
TWISTOR FUNCTIONS AND SHEAF COHOMOLOGY

Twistor Newsletter **2** (1976). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979; *Complex Manifold Techniques in Theoretical Physics.* Eds D. E. Lerner and P. D. Sommers, Pitman, 1979

Our group's discussions with Michael Atiyah, in which he lucidly explained to us about holomorphic sheaf cohomology, bore considerable fruit, as we began to understand the puzzling and fascinating aspects of twistor contour integrals in such terms. In this article I try to present the required basic ideas in a simple direct way. I have to point out, however, that my claim in the final section (*A finite covering for* $\overline{\mathbb{PT}}^+$) that the suggested pair of closed sets will suffice for what is needed in order to describe all massless wave functions, is in fact erroneous, as pointed out by Mike Eastwood; see Chapter 141. This erroneous final section was reprinted from Chapter 81.

Twistor Functions and Sheaf Cohomology. *Twistor Newsletter* **2** (1976). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979; *Complex Manifold Techniques in Theoretical Physics.* Eds D. E. Lerner and P. D. Sommers, Pitman, 1979

§2.2 TWISTOR FUNCTIONS AND SHEAF COHOMOLOGY *by R. Penrose*

Recall (Penrose 1975a,b) the following properties of a twistor function $f(Z^\alpha)$, to be used for generating a zero rest-mass field $\phi_{A'\ldots L'}$, or $\phi_{A\ldots K}$, by means of a contour integral,

$$\phi_{A'\ldots L'}(x) = (2\pi i)^{-2} \oint \rho_x \, \pi_{A'} \cdots \pi_{L'} \, f(Z^\alpha) \, d^2\pi,$$

$$\phi_{A\ldots K}(x) = (2\pi i)^{-2} \oint \rho_x \, \hat{\pi}_A \cdots \hat{\pi}_K \, f(Z^\alpha) \, d^2\pi:$$

(i) The function f is holomorphic on some domain $\mathcal{D}$ of twistor space *not* invariant under SU(2,2) (nor under the Poincaré group, nor the Lorentz group).

(ii) There is a "gauge" freedom G whereby

$$f \mapsto f + h^- + h^+,$$

where $h^\pm$ is holomorphic on some extended domain $\mathcal{D}^\pm (\supset \mathcal{D})$ of twistor space in which the contour γ can be deformed to a point (to the "left" in $\mathcal{D}^-$ and to the "right" in $\mathcal{D}^+$: see Fig. A).

(iii) G depends on the location of γ (and γ is not invariant under SU(2,2)--nor under the Poincaré group nor the Lorentz group--so G is not invariant either).

(iv) By invoking G, then moving γ, then invoking a new G, moving γ again, we can obtain a whole family of equivalent twistor functions, all giving the same field, the entire family being invariant under SU(2,2).

(v) This seems a little nebulous, and, for example, how do we add two such families (to give $\phi_{A'\ldots L'} + \psi_{A'\ldots L'}$), etc. etc?

Twistor Functions and Sheaf Cohomology. *Twistor Newsletter* **2** (1976). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979; *Complex Manifold Techniques in Theoretical Physics.* Eds D. E. Lerner and P. D. Sommers, Pitman, 1979

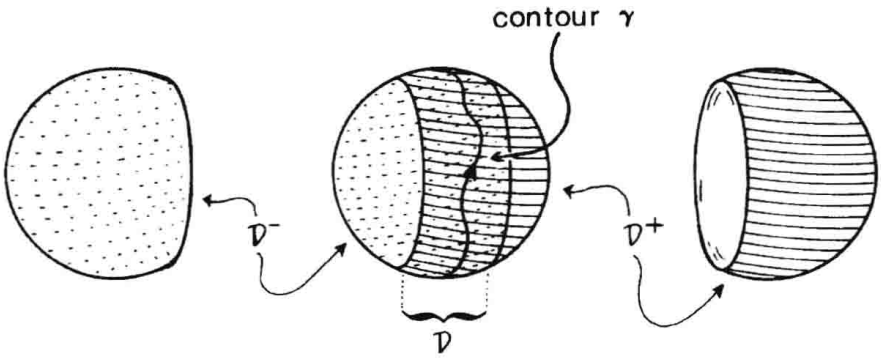


Figure A

A new viewpoint concerning twistor functions has been gradually emerging which makes good mathematical sense of all this: a twistor function is really to be viewed as a representative function (or cocycle) defining an element of a sheaf cohomology group. Now, the twistor theorist, when attacked by a purist for shoddiness in the domains, can counter-attack armed with his sheaf!

Thumbnail sketch of (relevant) sheaf cohomology theory

First let us recall how ordinary (Čech) cohomology works. Let X be a space (a Hausdorff paracompact topological space, say). Cover X with a locally finite system of open sets U_i. We define a *cochain* (with respect to this covering) with coefficients in an additive abelian group G (say, the integers Z, the reals R, or the complex field C) in terms of a collection of elements $f_i, f_{ij}, f_{ijk}, \ldots \in G$, assigned to the various U_i and their nonempty intersections: f_i assigned to U_i; f_{ij} assigned to $U_i \cap U_j$; f_{ijk} assigned to $U_i \cap U_j \cap U_k$; ... and $f_{ij} = -f_{ji}$, $f_{ijk} = -f_{jik} = f_{jki} = \ldots, \ldots$, i.e.

Twistor Functions and Sheaf Cohomology. *Twistor Newsletter* **2** (1976). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979; *Complex Manifold Techniques in Theoretical Physics.* Eds D. E. Lerner and P. D. Sommers, Pitman, 1979

$f_{i\ldots\ell} = f_{[i\ldots\ell]}$. Then

0-cochain $\alpha = (f_1, f_2, f_3, \ldots)$

1-cochain $\beta = (f_{12}, f_{23}, f_{13}, \ldots)$

2-cochain $\gamma = (f_{123}, f_{124}, \ldots)$

(where $U_1 \cap U_2$, $U_2 \cap U_3$, $U_1 \cap U_3, \ldots$, $U_1 \cap U_2 \cap U_3$, $U_1 \cap U_2 \cap U_4, \ldots$ are the non-empty intersections of U_i's).

Define *coboundary operator* δ as follows:

$$\delta\alpha = (\underbrace{f_2 - f_1}_{"f_{12}"}, \underbrace{f_3 - f_2}_{"f_{23}"}, \underbrace{f_3 - f_1}_{"f_{13}"}, \ldots)$$

$$\delta\beta = (\underbrace{f_{12} - f_{13} + f_{23}}_{"f_{123}"}, \underbrace{f_{12} - f_{14} + f_{24}}_{"f_{124}"}, \ldots)$$

etc.

(where, again, $U_1 \cap U_2, \ldots$ are the non-empty intersections). Then we have $\delta^2 = 0$. We call γ a *cocycle* if $\delta\gamma = 0$; we call γ a *coboundary* if $\gamma = \delta\beta$ for some β. Define the p^{th} *cohomology group* by

$$H^p_{\{U_i\}}(X,G) = \left(\begin{array}{c}\text{additive group}\\ \text{of p-cocycles}\end{array}\right) / \left(\begin{array}{c}\text{additive group}\\ \text{of p-coboundaries}\end{array}\right).$$

Note: $H^p_{\{U_i\}}(X,G)$, as defined, depends on the covering $\{U_i\}$. What we should do, to define $H^p(X,G)$, is to take the appropriate "limit" of all these $H^p_{\{U_i\}}(X,G)$ for finer and finer coverings $\{U_i\}$ of X. However (for X suitably non-pathological) we can always settle on a particular "sufficiently fine" covering $\{U_i\}$ where, in effect, there is no "relevant topology" left on each U_i or intersections thereof (i.e. all the $H^p(U_i \cap \ldots \cap U_k, G)$ vanish for all $p > 0$--although, as it stands, this is somewhat unhelpful because

a direct limit is then already involved in the definition of "sufficiently fine"). Then *this* $H^p_{\{U_i\}}(X,G) = H^p(X,G)$. I shall henceforth assume that such a "sufficiently" fine covering has been taken, and that it is countable and locally finite.

Now what does this definition have to do with the familiar "dual" relation to ordinary homology $H_p(X,G)$? How does γ assign values (elements of G) in a linear fashion to p-cycles in X, where γ is some element of $H^p(X,G)$? Intuitively, one adds together the f_{ij} for the shaded regions entered by the 1-cycle κ: see Fig. B. If γ is defined by $(f_{12}, f_{23}, \ldots)$, then $\gamma(\kappa) = f_{42} + f_{23} + f_{31} + \ldots$; the right hand side is a sum of (correctly signed) f's on regions entered by κ. The definition for higher dimensional p-cycles is similar (a complete rigorous discussion is easy enough).

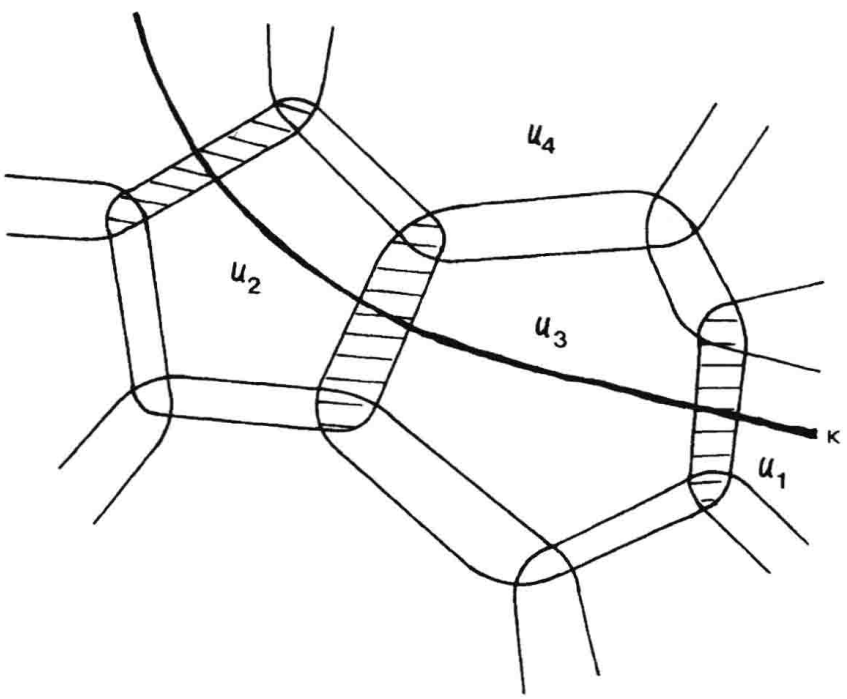


Figure B

Twistor Functions and Sheaf Cohomology. *Twistor Newsletter* **2** (1976). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979; *Complex Manifold Techniques in Theoretical Physics.* Eds D. E. Lerner and P. D. Sommers, Pitman, 1979

What about *sheaf* cohomology though? It's really a rather natural generalization of the above. But first we must rephrase the concept of a cochain slightly. Rather than thinking of f_i as simply an element of G "assigned" to U_i, and f_{ij} "assigned" to U_{ij}, etc., we think of f_i as a *function* defined on U_i which happens, in the above, to take this *constant* value "f_i" $\in$ G, and we think of f_{ij} as a constant function on $U_i \cap U_j$ with values in G, etc. (This has the incidental advantage that the requirement $U_i \cap U_j \neq \emptyset$ is now unnecessary.) This is still ordinary (Čech) cohomology. But the generalization to sheaf cohomology is now easily made: the functions $f_i, f_{ij}, f_{ijk}, \ldots$ are now *not* required to be constant. (In fact, we could even allow that the additive group G, in which the values of the f's reside, may vary from point to point in X, or, indeed, the f's need not really be "functions" in the ordinary sense at all--but such situations will not be considered here.) We may require that the f's be restricted in some way--in particular, for the purposes of the present applications we shall often require the f's to be *holomorphic* (with, here, G = C) and X a complex manifold--or, we may consider other related classes of functions.

So, what's a sheaf? Actually, I shan't even bother with a formal definition [which can be found in, for example, Morrow & Kodaira (1971); Gunning & Rossi (1965); Wells (1973)]. The essential point is that a sheaf is so defined that the Čech cohomology works just as well as before. In fact, a sheaf $\mathcal{S}$ defines an additive group G_U for each open set $U \subset X$. For example, G_U might be the additive group of all holomorphic functions on U (taking X to be a complex manifold). In this case we get the sheaf, denoted $\mathcal{O}$, of *germs of holomorphic functions* on X. Slightly more generally, we might consider "twisted" holomorphic functions, i.e. functions whose values are not just ordinary complex numbers, but taken in some complex line bundle

over X (think of "spin-weighted" functions, for example). An important example of such a twisted function would arise if X were taken to be *projective* twistor space PT (or a suitable portion thereof) and the functions considered were to be *homogeneous* (and holomorphic) of some fixed degree n in the twistor variable. For each open set $U \subset X$ we take G_U to consist of all such twisted functions on U, and the resulting sheaf, denoted $\mathcal{O}(n)$, is called the "sheaf of germs of holomorphic functions twisted by n" (on X). More generally we might consider functions whose values lie in some vector bundle B over X (e.g. we might consider tensor fields on X) and G_U would consist of the cross-sections of the portion of B lying above U.

Cochains are defined as before (with $f_i \in G_{U_i}$, $f_{ij} \in G_{U_i \cap U_j}$,...) and the coboundary operator δ, just as before. Then we obtain the p^{th} cohomology group of X, with coefficients in the sheaf S, as

$$H^p(X,S) = \left(\begin{array}{c}\text{p-cochains with}\\ \text{coefficients in } S\end{array}\right) \Big/ \left(\begin{array}{c}\text{p-coboundaries with}\\ \text{coefficients in } S\end{array}\right).$$

As before, we would need to take the appropriate limit for finer and finer coverings $\{U_i\}$ of X, but we can settle on one "sufficiently fine" covering if desired. Provided S is what is called a *coherent analytic* sheaf (and we are interested primarily in such sheaves—locally defined by n holomorphic functions factored out, if desired, by a set of s holomorphic relations), then "sufficiently fine" can be taken to mean that each of U_i, $U_i \cap U_j$, $U_i \cap U_j \cap U_k$,... is a *Stein manifold* (Gunning & Rossi 1965) (and it is sufficient just to specify that each U_i is Stein). In effect, a Stein manifold is a holomorphically convex open subset of C^n (or a domain of holomorphy). If X is Stein and S coherent, then $H^p(X,S) = 0$, if $p > 0$. Note: $\mathcal{O}$ and $\mathcal{O}(n)$ *are* coherent.

Twistor Functions and Sheaf Cohomology. *Twistor Newsletter* **2** (1976). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979; *Complex Manifold Techniques in Theoretical Physics.* Eds D. E. Lerner and P. D. Sommers, Pitman, 1979

Twistor functions as elements of $H^1(X,\mathcal{O}(n))$,

Let X be some suitable portion of projective twistor space PT , say some neighbourhood of a line in PT (corresponding to some neighbourhood of a point in Minkowski space), or say PT^+ , or PT^- . Suppose we can cover X with two sets U_1, U_2 (each open in X) such that every projective line L in X meets $U_1 \cap U_2$ in an annular region and where $U_1 \cap U_2$ corresponds to the domain of definition of some twistor function $f(Z^\alpha)$, homogeneous of degree n in the twistor Z^α (see Fig. C).

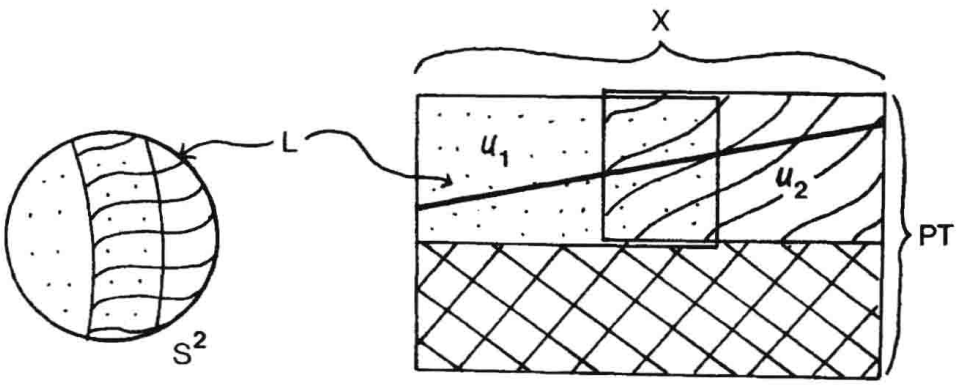


Figure C

Then $f = f_{12}$ is a twisted function on $U_1 \cap U_2$. There are no other $U_i \cap U_j$'s, so f_{12} by itself defines a 1-cochain β, with coefficients in $\mathcal{O}(n)$, for X. Clearly $\delta\beta = 0$, so β is a cocycle. The 1-coboundaries, for this covering, are functions of the form $h_2 - h_1$, where h_2 is holomorphic on U_2, and h_1 on U_1. Calling $\mathcal{D} = U_i \cap U_2$, $\mathcal{D}^- = U_1$, $\mathcal{D}^+ = U_2$, $h^- = -h_1$, and $h^+ = h_2$, we observe that the "equivalence" between twistor functions under the "gauge" freedom G that we started out with is just the normal cohomo-

logical equivalence between 1-cochains β,β' that their difference be a coboundary: $\beta' - \beta = \delta\alpha$, with $\alpha = (h_1,h_2)$. This suggests that we view the twistor function f as really defining us an element of $H^1(X,\mathcal{O}(n))$.

But this is all with respect to a particular covering of X, namely by $\{\mathcal{U}_1,\mathcal{U}_2\}$. Is this covering "fine" enough? There actually is a technical problem here. We cannot (normally) arrange that $\mathcal{U}_1$ and $\mathcal{U}_2$ are Stein manifolds (and in those exceptional cases when we can so arrange this, we would lose the invariance properties that we are striving for). It turns out, in fact, that this problem is not serious. One can show by direct construction (using the inverse twistor function--in the cases $n \leq -2$ at least, but probably in all cases) that for any *given* (analytic, positive frequency) field such a covering by two such sets $\mathcal{U}_1,\mathcal{U}_2$ is sufficient in the case $X = \overline{PT^+}$. Note that though X is invariant under SU(2,2), in this case, the *covering* is not. However, the cohomology group $H^1(X,\mathcal{O}(n))$ *is* invariant. Let us illustrate this by adding two elements of $H^1(X,\mathcal{O}(n))$ one of which is defined by a twistor function f, with respect to the covering $\{\mathcal{U}_1,\mathcal{U}_2\}$, and the other by $\hat{f}$, with respect to $\{\hat{\mathcal{U}}_1,\hat{\mathcal{U}}_2\}$, the second covering being a rotated version of the first (see Fig. D).

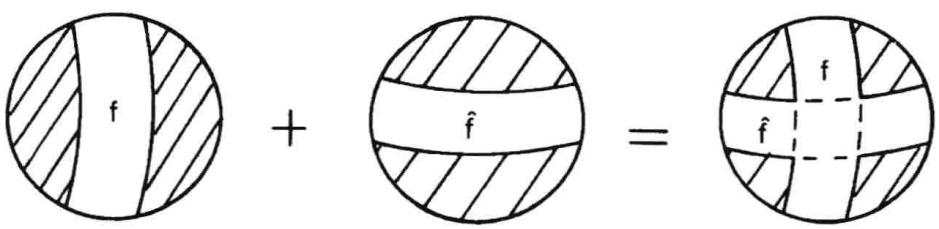


Figure D

Twistor Functions and Sheaf Cohomology. *Twistor Newsletter* **2** (1976). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979; *Complex Manifold Techniques in Theoretical Physics.* Eds D. E. Lerner and P. D. Sommers, Pitman, 1979

We define a representative cochain for the sum by taking the common refinement of both coverings. (Denote $\hat{\hat{U}}_1 = U_1 \cap \hat{U}_1, \hat{\hat{U}}_2 = U_1 \cap \hat{U}_2, \hat{\hat{U}}_3 = U_2 \cap \hat{U}_1$, $\hat{\hat{U}}_4 = U_2 \cap \hat{U}_2$ to give the refined covering $\{\hat{\hat{U}}_1, \hat{\hat{U}}_2, \hat{\hat{U}}_3, \hat{\hat{U}}_4\}$.) The 1-cocycle $\beta + \hat{\beta} = \hat{\hat{\beta}}$ is

$$(\hat{\hat{f}}_{12}, \hat{\hat{f}}_{13}, \hat{\hat{f}}_{23}, \hat{\hat{f}}_{14}, \hat{\hat{f}}_{24}, \hat{\hat{f}}_{34}) = (\hat{f}, f, f-\hat{f}, f+\hat{f}, f, \hat{f}).$$

Because of the "direct construction" argument mentioned above, this 1-cocycle will be cohomologous to (i.e. differing by a coboundary from) a cocycle of the form $(0, \hat{\hat{f}}, \hat{\hat{f}}, \hat{\hat{f}}, \hat{\hat{f}}, 0)$, so we can refer it back to the original covering if desired (although with U_1 and U_2 perhaps reduced slightly in size).

But we need not do so if we prefer not to. We have a generalization of the concept of a twistor function, namely as a collection of functions on portions of twistor space, defining a 1-cocycle with respect to some covering. We can actually use such a cocycle *directly*, obtaining the required space-time field by means of a *branched contour integral*. I shall just illustrate this with an example. Suppose that X is covered by 3 open sets U_1, U_2, U_3, where on a projective line L in X we get the picture of Fig.E.

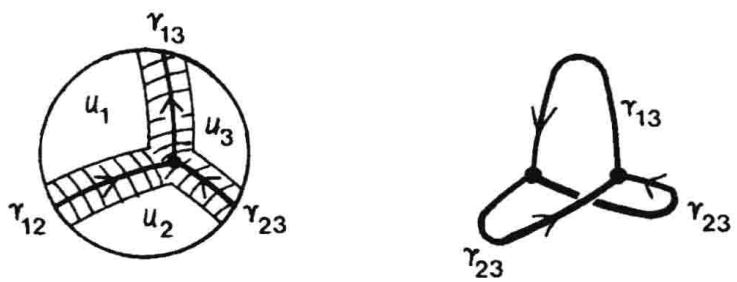


Figure E

It is a simple matter to check that the cocycle condition $f_{12} - f_{13} + f_{23} = 0$ (in $U_1 \cap U_2 \cap U_3$) ensures that the contour's end-points can be moved without affecting the result. This easily generalizes to coverings with N open sets. The 1-cocycle β is (f_{12}, f_{13}, f_{23}). To get the field $\phi....$ we perform the sum of three contour integrals with common end-points in $U_1 \cap U_2 \cap U_3$ (see Fig.E):

$$(2\pi i)^2 \phi.... = \int_{\Upsilon_{12}} ....f_{12} d^2Z + \int_{\Upsilon_{13}} ....f_{13} d^2Z + \int_{\Upsilon_{23}} ....f_{23} d^2Z.$$

Charge integrality in the "twisted photon"

We have (Morrow & Kodaira 1971) the exact sequence

$$0 \to Z \xrightarrow{\times 2\pi i} \mathcal{O} \xrightarrow{\exp} \mathcal{O}^* \to 0,$$

where $\mathcal{O}^*$ refers to non-zero holomorphic functions taken multiplicatively, from which we derive the long exact sequence:

$$... \to H^1(X,Z) \to H^1(X,\mathcal{O}) \to H^1(X,\mathcal{O}^*) \to H^2(X,Z) \to ...$$

ordinary integer 1st cohomology	$\phi_{AB} = \oint \hat{\pi}_A \hat{\pi}_B f\Delta$	"twisted photon"	ordinary integer 2nd cohomology

Choose X as a region in PT corresponding to a small space-time tube surrounding a charge world-line.

Then the topology of X is $S^2 \times S^2 \times R^2$, so $H^1(X,Z) = 0$, and $H^2(X,Z) \simeq Z \oplus Z$. Thus the space $H^1(X,\mathcal{O}^*)$ effectively contains $H^1(X,\mathcal{O})$, and is strictly larger than $H^1(X,\mathcal{O})$ if the map to $Z \oplus Z$ is not simply to the zero element. The image in the first Z is always zero, but the image in the second Z is the value of the *charge*. (This much follows, for example, from examination of §3.3.) From this we see that the $H^1(X,\mathcal{O}^*)$ description only works if the

Twistor Functions and Sheaf Cohomology. *Twistor Newsletter* **2** (1976). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979; *Complex Manifold Techniques in Theoretical Physics.* Eds D. E. Lerner and P. D. Sommers, Pitman, 1979

charge has *integer* value (i.e. lies in Z) whereas the $H^1(X,\mathcal{O})$ description only works if the charge value is *zero*: the contour integral formula

$$\phi_{AB} = (2\pi i)^{-2} \oint \rho_X \; \hat{\pi}_A \, \hat{\pi}_B \; f(Z^\alpha) \; d^2\pi,$$

where f is homogeneous of degree zero, implies vanishing charge. In contrast to this, it may be remarked that there is no restriction on the charge value if the -4-homogeneity $f(Z^\alpha)$ description

$$\tilde{\phi}_{A'B'} = (2\pi i)^{-2} \oint \rho_X \; \pi_{A'} \, \pi_{B'} f(Z^\alpha) d^2\pi$$

is used, this corresponding to $H^1(X,\mathcal{O}(-4))$.

A finite covering for $\overline{PT^+}$

A difficulty arises, in connection with the sheaf cohomology approach if one tries to find a *fixed* covering of $\overline{PT^+}$ which is "sufficiently fine" for *all* elements of $H^1(\overline{PT^+}, \mathcal{O}(n))$ (i.e. for all analytic massless wave functions). This arises basically because PT^+ is not holomorphically (pseudo-) convex and admits no finite covering by Stein manifolds. Any covering of PT^+ by Stein manifolds must necessarily be infinite, with open sets crowding up against the boundary PN. This is reflected in the fact that while for any *fixed* analytic massless wave function, a twistor function may be found (by way of the inverse twistor function) whose domain is just $\mathcal{U}_1 \cap \mathcal{U}_2$ where $\mathcal{U}_1$ and $\mathcal{U}_2$ are open sets with $\mathcal{U}_1 \cup \mathcal{U}_2 = PT^+$, this *particular* covering can never suffice for *all* such wave functions simultaneously (which may be arbitrarily "rough" and for which the corresponding twistor function's domain may need to be arbitrarily narrow).

However, if we allow ourselves to use *closed* sets for our sheaf cohomology (and there seems no reason against this), then the covering of $\overline{PT^+}$ by the

two closed sets, given respectively by

$$|Z^0 + Z^2| \geq |Z^1 + Z^3|, \quad Z^\alpha \bar{Z}_\alpha \geq 0$$

and by

$$|Z^0 + Z^2| \leq |Z^1 + Z^3|, \quad Z^\alpha \bar{Z}_\alpha \geq 0$$

will suffice (as the inverse twistor function shows) for *all* such wave functions.

Acknowledgment

I am grateful to George Sparling, Andrew Hodges, and Nick Woodhouse for many helpful discussions and, most particularly, to Michael Atiyah for explaining sheaf cohomology theory so clearly to us and for many insightful remarks.

REFERENCES

Gunning, R.C. & Rossi, H. 1965 *Analytic Functions of Several Complex Variables*. Prentice-Hall.

Morrow, J. & Kodaira, K. 1971 *Complex Manifolds*. Holt, Rinehart & Winston.

Penrose, R. 1975a *Twistor Theory, its Aims and Achievements*. In: *Quantum Gravity*, eds. C.J. Isham, R. Penrose & D.W. Sciama, pp. 268-407. Clarendon: Oxford.

Penrose, R. 1975b *Twistors and Particles*. In: *Quantum Theory and the Structures of Time and Space*, eds. L. Castell, M. Drieschner & C.F. von Weizsäcker, pp. 129-145.

Wells, R.O., Jr. 1973 *Differential Analysis on Complex Manifolds*. Prentice-Hall.

78
HOW TO GROW (OR AT LEAST HALF GROW) NEW TWISTORS FROM OLD

Twistor Newsletter **2** (1976). Also published in *Further Advances in Twistor Theory*, Vol. 1. Eds L. J. Mason and L. P. Hughston, Longman, 1990

An odd attempt to understand where twistors might (mathematically) 'come from'.

How to grow (or at least half grow) new twistors from old. *Twistor Newsletter* **2** (1976). Also published in *Further Advances in Twistor Theory*, Vol. 1. Eds L. J. Mason and L. P. Hughston, Longman, 1990

§I.4.1a How to grow (or at least $\frac{1}{2}$ grow) new twistors from old *by R.Penrose* (TN 2, June 1976)

As an alternative to the Čech description of $H^p(X,\mathcal{O})$ (see §O.2.2) we can take the additive group of all $\bar\partial$-closed $(0,p)$-forms modules the additive group of all $\bar\partial$-exact $(0,p)$-forms (cf. Morrow & Kodaira). By a (q,p)-form is meant a $(q+p)$-form on X (not generally holomorphic) having q holomorphic differentials and p anti-holomorphic differentials:

$$\alpha_{i_1\cdots i_q j_1\cdots j_p} dz^{i_1}\wedge\cdots\wedge dz^{i_q}\wedge d\bar z^{j_1}\wedge\cdots\wedge d\bar z^{j_p} = \alpha.$$

Then

$$\bar\partial\alpha = \frac{\partial\alpha_{i_1\cdots i_q j_1\cdots j_p}}{\partial\bar z^k} d\bar z^k\wedge dz^{i_1}\wedge\cdots\wedge dz^{i_q}\wedge d\bar z^{j_1}\wedge\cdots\wedge d\bar z^{j_p}$$

is a $\bar\partial$-exact $(q,p+1)$-form whose vanishing is the condition for α to be $\bar\partial$-closed. (We have $\bar\partial^2=0$.) Note that the $\bar\partial$-closed $(0,0)$-forms are holomorphic functions. We can also 'twist' such forms or functions, as before.

Applying this to $H^1(X,\mathcal{O}(n))$ we obtain the result that the helicity $\frac{1}{2}(-n-2)$ massless fields can be described in terms of forms $g^\alpha(Z,\bar Z)d\bar Z_\alpha$ on X, modulo forms

$$\frac{\partial h(Z,\bar Z)}{\partial\bar Z_\alpha} d\bar Z_\alpha, \qquad \text{where } \frac{\partial g^\alpha}{\partial\bar Z_\beta} = \frac{\partial g^\beta}{\partial\bar Z_\alpha},$$

the functions g^α being homogeneous of degree n in Z^α, and of degree -1 in $\bar Z_\alpha$ (to balance the $d\bar Z_\alpha$). In order that the form $g^\alpha d\bar Z_\alpha$ be actually defined on the projective space (appropriately twisted) we require also $g^\alpha\bar Z_\alpha=0$. The massless field integral now becomes

$$\varphi_{A\cdots D} = \frac{1}{2\pi i}\oint \frac{\partial}{\partial\omega^A}\cdots\frac{\partial}{\partial\omega^D} g^\alpha d\bar Z_\alpha I_{\beta\gamma} Z^\gamma dZ^\beta,$$

or

$$\varphi_{A'\cdots D'} = \frac{1}{2\pi i}\oint \pi_{A'}\cdots\pi_{D'} g^\alpha d\bar Z_\alpha I_{\beta\gamma} Z^\gamma dZ^\beta.$$

There is presumably much freedom in the choice of g^α. I shall suppose that its singularities can be arranged so that the contour in the above integrals (which is initially S^2, i.e. over the entire projective line corresponding to the space-time point in question) may be deformed, *freeing* $\bar Z_\alpha$ from Z^α in the process, until W_α ($=\tilde Z_\alpha=$ 'freed' $\bar Z_\alpha$) passes through the line I. Writing $W_\alpha = I_{\alpha\beta}X^\beta$, we get

$$g^\gamma d\bar Z_\gamma\wedge I_{\alpha\beta} Z^\alpha dZ^\beta = g^\gamma I_{\gamma\delta} dX^\delta\wedge I_{\alpha\beta} Z^\alpha dZ^\beta$$

The function g can be taken to be holomorphic in Z^α and W_α, with $W_\alpha g^\alpha = 0$ and $\partial^{[\alpha} g^{\beta]} = 0$ where $\partial^\alpha = \partial/\partial W_\alpha$. Restricting to $W_\alpha = I_{\alpha\beta} X^\beta$, we get $g^\alpha X^\beta I_{\alpha\beta} = 0$, so that $g^\alpha I_{\alpha\beta} = q X^\alpha I_{\alpha\beta}$ (for some $q(Z, X^\alpha I_{\alpha\beta})$ homogeneous, degree $(n,-2)$). Thus, the massless field integral can be written:

$$\varphi\ldots = \frac{1}{(2\pi i)^2} \oint \pi \cdots q X^\alpha dX^\beta I_{\alpha\beta} \wedge Z^\gamma dZ^\delta I_{\gamma\delta},$$

which is of the standard form for a 2-twistor integral. Thus we have conjured a new twistor out of thin air! This new twistor X^α is still only $\frac{1}{2}$ grown, though, since it appears as $X^\alpha I_{\alpha\beta}$. This can be stated as $I^{\alpha\beta} \partial q/\partial X_\alpha = 0$—which implies the massless relation

$$Z^\alpha X^\beta I_{\alpha\beta} I^{\gamma\delta} \frac{\partial^2}{\partial Z^\gamma \partial X^\delta} q = 0.$$

In fact, the $\bar{\partial}$-closed property $\partial^{[\alpha} g^{\beta]} = 0$ now disappears on $W_\alpha = X^\beta I_{\alpha\beta}$, being incorporated in $I^{\alpha\beta} \partial q/\partial X_\alpha = 0$, $X^\alpha \partial q/\partial X^\alpha = -2q$ (homogeneity relation).

79
SOME REMARKS ON THE 2-TWISTOR DESCRIPTION OF ELECTROMAGNETIC AND LINEARIZED GRAVITATIONAL POTENTIALS

Twistor Newsletter **2** (1976). Also published in *Further Advances in Twistor Theory*, Vol 1. Eds L. J. Mason and L. P. Hughston, Longman, 1990

Some remarks on the 2-twistor description of electromagnetic and linearized gravitational potentials. *Twistor Newsletter* **2** (1976). Also published in *Further Advances in Twistor Theory*, Vol 1. Eds L. J. Mason and L. P. Hughston, Longman, 1990

§I.4.1b Some remarks on the 2-twistor description of electromagnetic and linearized gravitational potentials *by R.Penrose* (TN 2, June 1976).

CPT considerations, as applied to 2-twistor functions, suggest that the photon and graviton, when they sit in the 2-twistor scheme, ought perhaps to be described by $(-2,-2)$-homogeneous functions, $F(Z,X)$ and $G(Z,X)$ respectively, where

$$\Phi_{AA'} = \frac{1}{(2\pi i)^4}\oint (Z_{A'}\frac{\partial}{\partial Z^A} - X_{A'}\frac{\partial}{\partial X^A})F d^2Z \wedge d^2X$$

and

$$h_{AA'BB'} = \frac{1}{(2\pi i)^4}\oint (Z_{A'}X_{B'}\frac{\partial^2}{\partial Z^A \partial X^B})G d^2Z \wedge d^2X$$

We have the Lorenz $\nabla^a \Phi_a = 0$ and the DeDonder $\nabla^a(h_{ab} - \frac{1}{2}h^a{}_a g_{ab}) = 0$ gauges $(h_{ab} = h_{ba})$.

80
THE COSMIC CENSORSHIP HYPOTHESIS

Lecture delivered at International Astronomical Union, Warsaw (Copernicus), 1976

The Cosmic Censorship Hypothesis

by Roger Penrose
Mathematical Institute, Oxford.

The past eight years has seen remarkable changes in what has become generally acceptable by astronomers, from among the more bizarre predictions of the theoreticians. First the neutron star has arrived, with its mass of the order of the sun compressed to within a radius of some ten kilometres - and now probably the black hole, that disembodied gravitational field created by matter which has long since crushed itself out of existence by the strength of its own gravitational pull. Do there remain skeletons, wierder still, in the theoreticians' cupboards, which might someday have to be reckoned with by practical astronomers? Perhaps the most alarming possibility yet is the naked singularity. While not exactly a positive theoretical prediction, it remains one in the negative sense - that theoreticians have so far found no way of ruling it out!

To explain the difficulty more fully, let us recall the normal picture of spherical collapse to a black hole. This picture has two features of particular note, namely the absolute event horizon, that sphere within which no signal can escape to the outside world, and the space-time singularity at the centre, where curvatures become so large that some new physical theory is needed. The (as yet) unpredictable effects arising near the

singularity have no observational significance, since they are shielded by the horizon. But whether this necessarily remains true when asymmetries are present remains an open question. Collapse past a 'point of no return' can be characterized by the presence of some light cone, the rays of which begin to reconverge after sufficient matter falls across them. Then a singularity theorem can be invoked, proving on the basis of classical general relativity that a space-time singularity occurs. But is this singularity necessarily hidden from view, in which case we necessarily have a black hole, or may it sometimes be visible - a naked singularity? The hypothesis that naked singularities are forbidden in generic collapse is called cosmic censorship.

The results of perturbation analysis, and certain other arguments, indicate rather strongly that if the collapse is initially only slightly asymmetrical, then cosmic censorship holds. But little can yet be said with confidence concerning highly asymmetrical collapse. Various attempts at finding a contradiction with some implications of cosmic censorship have so far failed. But attempts at finding a direct proof of the hypothesis have met with even less success. Even stating the hypothesis in precise mathematical terms (something like 'naked singularities would be unstable') has proved elusive. The problem remains the most important unanswered question of classical general relativity.

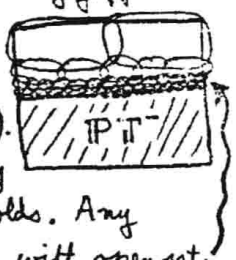
81
A FINITE COVERING FOR PT+

Twistor Newsletter **3** (1976)

The suggestion I make here (reprinted in Chapter 77) that one can obtain all the massless free-field wave functions from the pair of closed sets given here is *incorrect*, as was pointed out to me by Mike Eastwood; see Chapter 141. I had made an error in my application of the inverse twistor function operation (a genuine and useful procedure, described in Chapter 69, pp. [311]–[316]).

A Finite Covering for $\overline{\mathbb{P}\mathbb{T}^+}$

A difficulty arises, in connection with the sheaf cohomology approach to twistor functions (see T.N.2), if one tries to find a fixed covering of $\overline{\mathbb{P}\mathbb{T}^+}$ which is "sufficiently fine" for all elements of $H^1(\overline{\mathbb{P}\mathbb{T}^+}, \mathcal{O}(n))$ (i.e. for all analytic massless wave functions). This arises basically because $\mathbb{P}\mathbb{T}^+$ is not holomorphically (pseudo-) convex and admits no finite covering by Stein manifolds. Any covering of $\mathbb{P}\mathbb{T}^+$ by Stein manifolds must necessarily be infinite, with open sets crowding up against the boundary $\mathbb{P}\mathcal{N}$. This is reflected in the fact that while for any fixed analytic massless wave-function, a twistor function may be found (by way of the inverse twistor function) whose domain is just $\mathcal{U}_1 \cap \mathcal{U}_2$ where $\mathcal{U}_1$ and $\mathcal{U}_2$ are open sets with $\mathcal{U}_1 \cup \mathcal{U}_2 = \mathbb{P}\mathbb{T}^+$, this particular covering can never suffice for all such wave functions simultaneously (which may be arbitrarily "rough" and for which the corresponding twistor function's domain may need to be arbitrarily narrow).



However, if we allow ourselves to use closed sets for our sheaf cohomology (and there seems no reason against this), then the covering of $\overline{\mathbb{P}\mathbb{T}^+}$ by the two closed sets, given respectively by $|Z^0+Z^2| \geq |Z^1+Z^3|$, $Z^\alpha \bar{Z}_\alpha \geq 0$ and by $|Z^0+Z^2| \leq |Z^1+Z^3|$, $Z^\alpha \bar{Z}_\alpha \geq$ inverse twistor function shows) for all such wave functions.

R. Penrose

82
THE BACK-HANDED PHOTON (OR, TO A CRICKETER, THE GOOGLY PHOTON!)

Twistor Newsletter **3** (1976). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979; *Complex Manifold Techniques in Theoretical Physics.* Eds D. E. Lerner and P. D. Sommers, Pitman, 1979

This appears to be the first mention of the term 'googly' in twistor theory. The title given here is taken from the original *Twistor Newsletter* article but, by the time the article was reproduced in *Advances in Twistor Theory,* from where this reproduction comes, the term 'googly' had become established in twistor vocabulary.

90 The back-handed photon (or, to a cricketer, the **googly** photon!). *Twistor Newsletter* **3** (1976). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979; *Complex Manifold Techniques in Theoretical Physics.* Eds D. E. Lerner and P. D. Sommers, Pitman, 1979

§3.7. THE GOOGLY PHOTON* *by R. Penrose*

Current twistor dogma presents the following picture of things:

(i) The wave function of a particle or system or particles is to be given by a "twistor function" $\psi(X^\alpha,\ldots,Z^\alpha)$, which is holomorphic in all arguments $X^\alpha,\ldots,Z^\alpha$ --possibly with the modification that some (or all) of these should be dual twistors $W_\alpha,\ldots$ instead (or, conceivably, Hughston-type multi-twistors $X^{\alpha\beta\gamma},\ldots$, etc.).

(ii) More correctly, ψ is (presumably) an element of a *sheaf cohomology group*, say $\psi \in H^k(Q,\mathcal{O})$, for some suitable open set (or closure of an open set) $Q \subset T \times T \times \ldots \times T$ say (T = twistor space), $\mathcal{O}$ being the sheaf of (germs of) holomorphic functions on Q (see Chapter 2).

(iii) When a particle interacts it does so by *deforming* the structure of twistor space, or of $Q \subset T \times \ldots \times T$, whereby ψ now plays an *active* role in determining the precise deformation in question, thence affecting the behavior of other twistor functions defined on Q.

The above is vague in various respects. But something like (iii) is strongly suggested by two examples: the *non-linear graviton* (see §3.1) and the *twisted photon* (§3.2). (In fact, for gravitons, (iii) is strictly true only in the weak-field limit--unless some kind of non-linear sheaf theory can be evoked.) In the standard procedures for performing deformations (infinitesimal ones at least) it is only H^1's that are involved. Sparling has suggested that *new* procedures for performing deformations may be needed in order to be able to use H^k's ($k > 1$) in an active way. For the moment, I

---

* 'Googly' is a cricketing term, meaning a positive helicity cricket ball bowled with an apparently negative helicity bowling action.

prefer the more modest idea that *single* particles are to be described by H^1's, these being the things that interact *directly* with other particles. H^2's, H^3's, etc. would play roles in describing many-particle states, with elements of suitable $H^k(Q, \mathcal{O})$ groups defining k-particle states. This could supply a possible answer to a question posed to me by Feynman: how do you know whether $\psi(X^\alpha, Y^\alpha, Z^\alpha)$ describes a hadron, or three massless particles, or a lepton and a massless particle? Elements of $H^1(Q, \mathcal{O})$, $H^3(Q, \mathcal{O})$, and $H^2(Q, \mathcal{O})$ are really quite different kinds of animals, after all! But the detailed implementation of this idea has proved elusive--not the least problem being to understand "twistorially" why the spin-statistics relation should hold for many-particle states.

Perhaps the most primitive problem, in trying to push forward with (iii), lies in the fact that a twisted photon is only *half* a photon, namely the left-handed half. Thus, the twisted photon construction gives a deformation of $Q = T^+$ starting from a twistor function $f(Z^\alpha)$, homogeneous of degree zero [$f \in H^1(PT^+, \mathcal{O}(0))$], to give a *left*-handed photon. Of course a right-handed photon can be produced by using $\tilde{f}(W_\alpha)$, of degree zero. But such would be to defeat the purpose of the economy of the twistor description (i). Simply changing the helicity quantum number (or any other quantum number) should not involve us in changing the *space* Q over which the twistor function is defined. Thus maintaining the general programme (i) - (iii) seems to lead us to the view that some form of deformation of T^+ *must* be possible, which effectively encodes the information provided by an element $f \in H^1(T^+, \mathcal{O})$, when f is homogeneous of degree -4 [i.e., in effect, $f \in H^1(PT^+, \mathcal{O}(-4))$]. My suggestion for this, which relates closely to an earlier proposal due to Sparling and Ward (cf. §2.4), is as follows.

Consider, first, the standard Ward twisted photon. This is obtained by

The back-handed photon (or, to a cricketer, the **googly** photon!). *Twistor Newsletter* **3** (1976). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979; *Complex Manifold Techniques in Theoretical Physics.* Eds D. E. Lerner and P. D. Sommers, Pitman, 1979

deforming the bundle T^+ where the base space remains the undeformed PT^+ and the fibre $C - \{0\}$. Furthermore, the Euler operator $Z^\alpha \frac{\partial}{\partial Z^\alpha} =: T_Z$ remains globally defined. However, with non-trivial twisting, the forms $DZ: = \frac{1}{24} \varepsilon_{\alpha\beta\gamma\delta}\, dZ^\alpha \wedge dZ^\beta \wedge dZ^\gamma \wedge dZ^\delta$ and $\mathcal{D}Z: = \frac{1}{6} \varepsilon_{\alpha\beta\gamma\delta}\, Z^\alpha dZ^\beta \wedge dZ^\gamma \wedge dZ^\delta$ are *not* well defined; indeed, globally defined analogues of these forms do not exist at all on a non-trivially twisted Ward photon. There is a good reason for this. In the case of flat T, there is a canonical way of representing a twistor Z^α (up to the fourfold ambiguity $\pm Z^\alpha$, $\pm iZ^\alpha$) in terms of PT: namely as the projective twistor Z in PT, together with a *3-form at* Z in PT. This 3-form lifts into T along the fibre over Z; where it agrees with $\mathcal{D}Z$ defines us the point Z^α (or iZ^α, or $-Z^\alpha$, or $-iZ^\alpha$). Thus, if $\mathcal{D}Z$ is canonically known in the bundle, then the complete bundle structure is determined uniquely in terms of PT; therefore no twisting in the "photon" can occur. (This construction is the analogue of how one defines a spinor κ^A (up to sign) in terms of the celestial sphere.) Furthermore, if a $\mathcal{D}Z$ exists globally in a bundle over PT^+, then it must be unique to an overall constant factor.

Suppose, generally, we have a bundle (or holomorphic fibration) *locally*, which is just a complex 4-space over a complex 3-space with fibre a complex 1-space. Then all we know locally in the 4-space is a direction field δ_Z (1-foliation). To know $\mathcal{D}Z$ in the 4-space would be to know rather more structure than δ_Z. Being a 3-form, $\mathcal{D}Z$ is orthogonal to (i.e., annihilates) precisely one complex direction and so serves to *define* δ_Z. (I shall always assume $\mathcal{D}Z$ to be restricted to be orthogonal to δ_Z, so as to qualify to be a "$\mathcal{D}Z$.") Now $\mathcal{D}Z$ also defines a volume 4-form DZ by $d\mathcal{D}Z = 4DZ$. Furthermore $\mathcal{D}Z$ and DZ together define the Euler operator T_Z, roughly speaking by "$T_Z = \mathcal{D}Z \div DZ$," or more precisely by $\mathcal{D}Z = T_Z \lrcorner DZ$. Conversely, this relation

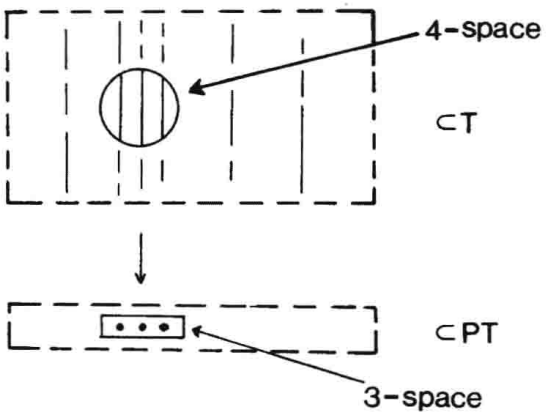


Figure A

shows us that $\mathcal{D}Z$ is determined by the *pair* (T_Z, DZ). To know one or the other of T_Z, DZ is, by itself, not sufficient to determine $\mathcal{D}Z$, but the two together are equivalent to $\mathcal{D}Z$. Furthermore, assuming that T_Z is to point along δ_Z (where δ_Z is given), there is precisely as much information in T_Z locally as there is in DZ. Each provides us with a kind of local scaling, but it is a "homogeneity degree 4" scaling for DZ and a "homogeneity degree 0" scaling for T_Z. In a sense DZ and T_Z seem to be sorts of duals to one another.

We can regard the twisted photon (§3.2) as arising when we retain only the T_Z scaling and throw out DZ. Let us try to do the "dual" thing and retain DZ while throwing out T_Z. I shall proceed in a fairly explicit way, assuming that two "coordinate" patches are given, where a standard twistor description is given in each patch, with X^α on the left and Z^α on the right. I am assuming there is no monkey-business in the base space, so $X^\alpha \propto Z^\alpha$ may be assumed on the overlap region. (The fibres in each half are given when X^α or Z^α is held constant up to proportionality.) The hypothesis is now

 The back-handed photon (or, to a cricketer, the **googly** photon!). *Twistor Newsletter* **3** (1976). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979; *Complex Manifold Techniques in Theoretical Physics.* Eds D. E. Lerner and P. D. Sommers, Pitman, 1979

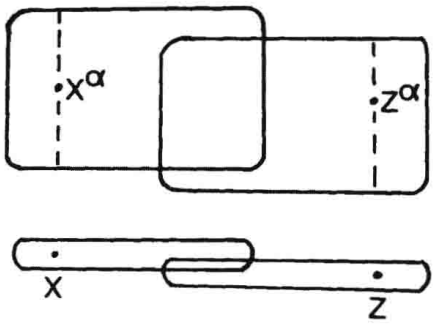


Figure B

DX = DZ on the overlap (instead of Ward's $T_X = T_Z$), i.e., $d\mathcal{D}X = d\mathcal{D}Z$, i.e., $d\{\mathcal{D}X-\mathcal{D}Z\} = 0$. But $\delta_X = \delta_Z$, so $\mathcal{D}X \propto \mathcal{D}Z$. Put $\mathcal{D}X = \{1 + f(Z^\alpha)\}\mathcal{D}Z$; then $d\{f(Z^\alpha)\mathcal{D}Z\} = 0$, which holds iff $f(Z^\alpha)$ is *homogeneous of degree -4.* The transition relation is then

$$X^\alpha = \{1 + f_{-4}(Z^\alpha)\}^{\frac{1}{4}}\, Z^\alpha. \qquad (1)$$

This appears to be a particular case of a Sparling-Ward construction given earlier for massless fields of *any* helicity.

We run into trouble owing to branch points arising from the fourth root unless we exclude an extended region about the origin of each fibre. But things are okay near ∞ on the fibres. Thus, I envisage the fibres as being like C with some bounded (probably connected) region removed (i.e., biholomorphic to $C - \{z : |z| < 1\}$). Note that T_Z is not preserved by the patching, but related by $T_X = (1 + f(Z^\alpha))T_Z$ (which does strange things to the notion of a homogeneous function).

The intention is to regard $f(Z^\alpha)$ as a twistor function for a right-handed photon. But, in accordance with (ii), such an $f(Z^\alpha)$ is really describing an element of a sheaf cohomology group. Thus we need to check that (1) is appropriately cohomological. Suppose we have a covering of the base space

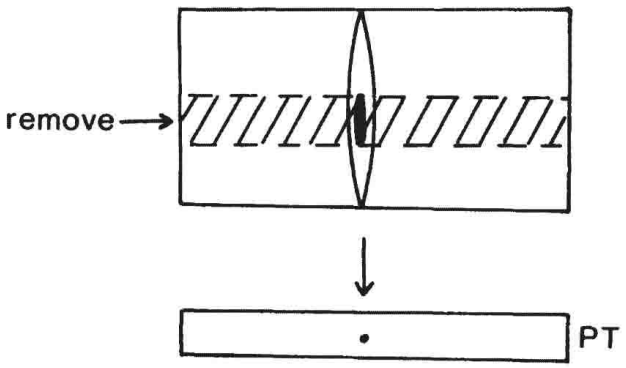


Figure C

(say PT^+) by a number of open sets U_i. Then on each $U_i \cap U_j$ we need $f_{ij} = -f_{ji}$ such that on each $U_i \cap U_j \cap U_k$, $f_{ik} = f_{ij} + f_{jk}$ (cocycle condition). If we use standard twistor "coordinates" Z^α for the entire base space, then the f_{ij} will simply be homogeneous functions of Z^α, of degree -4. Thus, $f_{ij}(Z^\alpha) = -f_{ji}(Z^\alpha)$, $f_{ik}(Z^\alpha) = f_{ij}(Z^\alpha) + f_{jk}(Z^\alpha)$. The cocycle defining the right-handed photon can be taken to be just this collection of homogeneous functions. We piece together our "bundle" by taking "coordinates" Z^α_i for the portion lying above U_i, where on each overlap (say, the portion of the bundle lying above $U_i \cap U_j$) we have $Z^\alpha_i = \{1 + f_{ij}(Z^\beta_j)\}^{\frac{1}{4}} Z^\alpha_j$. We must first check the consistency of this with $f_{ij} = -f_{ji}$. Avoiding messy indices we have (1) : $X^\alpha = \{1 + f(Z^\beta)\}^{\frac{1}{4}} Z^\alpha$ and $Z^\alpha = \{1 + g(X^\beta)\}^{\frac{1}{4}} X^\alpha$, and we need to check that $g = -f$. Now $Z^\alpha = \{1 + g(X^\beta)\}^{\frac{1}{4}}\{1 + f(Z^\beta)\}^{\frac{1}{4}} Z^\alpha$, so we require $(1 + g(X^\beta))(1 + f(Z^\beta)) = 1$; that is

$$1 = (1 + g(\{1 + f(Z^\beta)\}^{\frac{1}{4}} Z^\sigma))\ (1 + f(Z^\rho))$$
$$= (1 + \frac{g(Z^\sigma)}{1 + f(Z^\beta)})(1 + f(Z^\rho))$$

 The back-handed photon (or, to a cricketer, the **googly** photon!). *Twistor Newsletter* **3** (1976). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979; *Complex Manifold Techniques in Theoretical Physics.* Eds D. E. Lerner and P. D. Sommers, Pitman, 1979

$$= (1 + f(Z^\rho) + g(Z^\rho))$$

(because of the -4 homogeneity of g) so $g(Z^\rho) = -f(Z^\rho)$ as required. Next we check the compatibility of (1) with $f_{ik} = f_{ij} + f_{jk}$. Again avoiding indices we have, on the triple intersection region about $U_i \cap U_j \cap U_k$, relations

$$Y^\alpha = \{1 + p(X^\rho)\}^{\frac{1}{4}}X^\alpha;\ X^\rho = \{1 + f(Z^\sigma)\}^{\frac{1}{4}}Z^\rho;\ Y^\alpha = \{1 + q(Z^\sigma)\}^{\frac{1}{4}}Z^\alpha\ .$$

For consistency we require

$$\{1 + p(\{1 + f(Z^\sigma)\}^{\frac{1}{4}}Z^\rho\})\{1 + f(Z^\beta)\} = 1 + q(Z^\sigma)$$

$$\text{i.e.}\quad \{1 + \frac{p(Z^\rho)}{1 + f(Z^\rho)}\}\ \{1 + f(Z^\rho)\} = 1 + q(Z^\rho)$$

i.e. $q(Z^\rho) = p(Z^\rho) + f(Z^\rho)$ as required. We need also to show that if the cocycle is a coboundary, i.e., if $f_{ij} = f_j - f_i$ for each i,j, then the bundle is the same as that for flat twistor space (f_i being holomorphic throughout U_i). For this we set $Z^\alpha_i = \{1 + f_i(Z^\beta_i)\}^{\frac{1}{4}}Z^\alpha_i$ and find that the Z^α_i, so constructed, is defined *globally* over the whole bundle. The compatibility over each $U_i \cap U_j$ follows by a calculation which is basically the same as the ones given above. Finally we need the fact that *if* the bundle is the same as that for flat twistor space (where here and above "the same" must be suitably interpreted in terms of "analytically extendible to"--owing to the gaping holes in the fibres), then the $\{f_{ij}\}$-cocycle is a coboundary. To prove this, we simply reverse the above argument. The flatness implies the existence of a global Z^α for the bundle. The local Z^α_i for each patch (above U_i) must be related to Z^α by a formula $Z^\alpha_i = \{1 + f_i(Z^\beta_i)\}^{\frac{1}{4}}Z^\alpha_i$ from which follows (by the reverse of the above calculation)

$f_{ij} = f_j - f_i$ above $U_i \cap U_j$.

The argument just given effectively shows that although (1) appears to be a highly non-linear relation, the system of bundles so constructed has nevertheless a *linear* structure (given by simply adding the sheaf cohomology group elements). This may be contrasted with Sparling's method of patching together non-linear gravitons:

$$X^\alpha = \exp\{I^{\rho\sigma} \frac{\partial f(Z^\tau)}{\partial Z^\rho} \frac{\partial}{\partial Z^\sigma}\} Z^\alpha.$$

For this, the cocycle condition *fails*. This is a manifestation of the very non-linearity of the non-linear graviton. On the other hand, the "4" in the above construction is *not* essential for linearity and could be replaced by other powers; i.e., f of homogeneity -n, with $X^\alpha = \{1 + f(Z^\beta)\}^{1/n} Z^\alpha$. When n = 1, "0", -1, -2, -3,... this is implicit in the Sparling-Ward construction. However, when $n < 0$ the global structure of the fibres would have to be something different (because $(1 + z^{-n})^{1/n} z \to 1$ as $|z| \to \infty$ if $n < 0$, whereas $(1 + z^{-n})^{1/n} z \to z$ as $|z| \to \infty$ if $n > 0$). In any case, the motivation from "preservation of DZ" exists only when n = 4.

A direct construction of the right-handed Maxwell field $\tilde{\phi}_{A'B'}$ (satisfying $\nabla^{AA'} \tilde{\phi}_{A'B'} = 0$) from the bundle structure has not yet emerged. But the most serious problem is that of fitting the right-handed and left-handed ways of deforming T^+ together into one bundle...? One possibility for doing this is a method suggested by Hughston in §3.8. However it is unclear, as yet, how to extract left- and right-handed Maxwell fields which do not interact with one another. If this problem can be resolved, then among other things, it *might* suggest an analogous approach to an ambidextrous graviton!

83
ANY SPACE-TIME HAS A PLANE WAVE AS A LIMIT

Differential Geometry and Relativity. Eds M. Cahen and M. Flato, pp. 271–5. *Mathematical Phys. and Appl. Math.*, Vol. 3, Reidel, Dordrecht, 1976

This limiting procedure appears to have interest in string theory and related areas, and is referred to as the 'Penrose limit'.

100 Any space-time has a plane wave as a limit; *Differential Geometry and Relativity*. Eds M. Cahen and M. Flato, pp. 271–5. *Mathematical Phys. and Appl. Math.*, Vol. 3, Reidel, Dordrecht, 1976

ROGER PENROSE

ANY SPACE-TIME HAS A PLANE WAVE AS A LIMIT

It is a great pleasure for me to honour André Lichnerowicz on his 60th birthday. And since he has done so much to make relativity respectable as a branch of mathematics, I feel it is appropriate here to point out a simple, yet perhaps surprising, general mathematical property of hyperbolic normal pseudo-Riemannian manifolds which is motivated from knowledge of certain well-known exact solutions in general relativity.

The concept of tangent space at a point p in a manifold M is, of course, basic to differential geometry. Intuitively, one may envisage smaller and smaller neighbourhoods of p in M which are correspondingly scaled up by larger and larger factors. In the limit we obtain the tangent space T_p to M at p. However, it does not appear to have been noted in the literature before that a corresponding procedure can also be applied to any entire (properly embedded) *null geodesic* γ in any space-time or, more generally, in any hyperbolic normal pseudo-Riemannian manifold. But whereas T_p is, in an essential way, a flat space, the corresponding procedure applied to γ yields a curved space W_γ known as a *plane wave*.

I shall give the argument explicitly for a four-dimensional space-time, the generalization to n-dimensions being obvious. It is well-known (cf. [1], for example) that a coordinate system can be set up in some neighbourhood Q of a conjugate point-free portion γ', of γ, the metric components taking the form

$$g_{ij} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & a & b_2 & b_3 \\ 0 & b_2 & -c_{22} & -c_{23} \\ 0 & b_3 & -c_{23} & -c_{33} \end{pmatrix}, \tag{1}$$

where $a, b_2, \ldots, c_{33}$ are smooth functions of the coordinates x^0, x^1, x^2, x^3, the c's forming a positive-definite matrix. (To set up such a coordinate system, we choose a null hypersurface through γ' belonging to a smooth family of null hypersurfaces labelled by $x^1 = \text{const}$. The null geodesic generators of these hypersurfaces are to be given by

Cahen & Flato (eds.), Differential Geometry and Relativity, 271–275. *All Rights Reserved. Copyright © 1976 by D. Reidel Publishing Company, Dordrecht-Holland.*

x^1, x^2, $x^3 = \text{const.}$, x^0 being an affine parameter on each one. In particular, γ' is to be given by $x^1 = x^2 = x^3 = 0$.)

Consider next, a new coordinate system (X^0, X^1, X^2, X^3) for Q

$$X^0 = x^0, \qquad \Omega^2 X^1 = x^1, \qquad \Omega X^2 = x^2, \qquad \Omega X^3 = x^3, \tag{2}$$

where Ω is a positive number, and define

$$g_{ij}(\Omega) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & \Omega^2 A & \Omega B_2 & \Omega B_3 \\ 0 & \Omega B_2 & -C_{22} & -C_{23} \\ 0 & \Omega B_3 & -C_{23} & -C_{33} \end{pmatrix}, \tag{3}$$

where $A(X^0, X^1, X^2, X^3) = a(x^0, x^1, x^2, x^3)$, etc. Then

$$g_{ij}(\Omega)\, \mathrm{d}X^i\, \mathrm{d}X^j = \Omega^{-2} g_{ij}\, \mathrm{d}x^i\, \mathrm{d}x^j. \tag{4}$$

Thus the metric components (3), in the X^i-coordinate system, define a metric $g(\Omega)$ which is a multiple of (i.e. conformal to, with constant scale factor) the original space-time metric g in Q. Now let $\Omega \to 0$. We obtain, for the limiting version of (4)

$$2\, \mathrm{d}X^0\, \mathrm{d}X^1 - C_{22}(\mathrm{d}X^2)^2 - 2C_{23}\, \mathrm{d}X^2\, \mathrm{d}X^3 - C_{33}(\mathrm{d}X^3)^2, \tag{5}$$

where, because of relations (2), the coefficients C_{22}, C_{23}, C_{33} can depend on the *one* variable X^0 only. The metric form (5) is the familiar Rosen expression [2] for a plane wave – a space-time with a five-parameter group of isometries [3].

What has been established is a limiting procedure whereby the immediate neighbourhood of γ' is expanded outwards, the metric being scaled uniformly up at the same time, and an exact plane wave W_γ is the result. If desired, we may view this limiting procedure in terms of a five-dimensional manifold-with-boundary with coordinates $\Omega, X^0, X^1, X^2, X^3$, the boundary being given by $\Omega = 0$ (cf. Geroch [4]). This boundary $\Omega = 0$ has the intrinsic metric of a plane wave, and for each $\Omega = \text{const.} > 0$ we get a space whose metric is just Ω^{-2} times that of M in Q.

The reason that the metric must be scaled up is that there is a 'boost' involved in the coordinate change which, as Ω gets smaller, corresponds to a velocity approaching the limiting value unity. A general curvature tensor will appear to 'blow up' as the coordinates (or tetrad

frame) are boosted to the speed of light* (cf. [5]). So, for a finite answer, we must correspondingly scale the metric up, whence the curvature is scaled down.

There is a 'physical' interpretation of the above mathematical procedure, which is the following. We envisage a succession of observers travelling in the space-time M whose world lines approach the null geodesic γ more and more closely; so we picture these observers as travelling with greater and greater speeds, approaching that of light. As their speeds increase they must correspondingly recalibrate their clocks to run faster and faster (assuming that all space-time measurements are referred to clock measurements in the standard way [7]), so that in the limit the clocks measure the affine parameter x^0 along γ. (Without clock recalibration a degenerate space-time metric would result.) In the limit, the observers measure the space-time to have the plane-wave structure W_γ.

In view of this universal role that the plane waves have in relation to general space-times, it is worth recalling a certain peculiar feature that they possess. Throughout the history of the subject, this feature has, in its various guises, often resulted in a belief that plane waves are in some special way 'non-physical' even as idealizations of the kind that one would expect plane waves to be. For example, Rosen [2] pointed out that any non-flat vacuum metric of the form (5) necessarily encounters singularities when one attempts to extend the range of the coordinates in order to obtain a complete space-time. This led to the erroneous belief, for many years, that gravitational plane waves did not exist in general relativity. However, it was later shown by Robinson [8] (cf. Bondi *et al.* [3]) that this was merely a coordinate effect† and that a complete plane wave could be covered by using a system of overlapping coordinate neighbourhoods, in each of which the metric takes the form (5). Another way in which this peculiarity of plane waves exhibits itself

* The Weyl tensor of a generic space-time approaches a *null* type (all principal null directions coincident) when a high velocity 'boost' is applied in a general direction [5, 6]. This is consistent with the fact that all plane waves have null (or else zero) Weyl tensors. Pirani [6] established this result in the case of the Schwarzschild solution only. But as we see from the above, the phenomenon is quite general; moreover, the entire metric of a smooth plane wave arises, not just its curvature tensor at one point.

† Robinson also showed that a coordinate system for the whole of W_γ can be given for which the metric takes the form $\mathrm{d}u\,\mathrm{d}v + h_{ij}(u)x^i x^j\,\mathrm{d}u^2 - \mathrm{d}x^i\,\mathrm{d}x^i$. This form of metric is due originally to Brinkman [9] who had obtained it as a special case of a more general class of vacuum metrics known as plane-fronted waves.

is in the non-existence of a global Cauchy hypersurface and the consequent property of non-embeddability in hyperbolic normal flat spaces of arbitrarily large dimension (Penrose [10]).

In the present context this phenomenon arises because of the possible existence of pairs of conjugate points on γ [1, 11]. Coordinates for which the metric tensor components take the form (1) cannot cover a portion of γ containing a pair of conjugate points (cf. [1]). Thus, for the whole of γ, a system of overlapping coordinate neighbourhoods would be required whenever γ contains pairs of conjugate points. Furthermore [12], provided M satisfies the weak energy condition, null geodesic completeness and a (physically reasonable) generality condition, it follows that *every* null geodesic actually does contain a pair of conjugate points. Thus, the peculiar features of plane waves mentioned above are in some limiting sense already present in every such physically reasonable space-time. (This does not, however, prevent such space-times from having global coordinate systems or Cauchy hypersurfaces, or from being embeddable in hyperbolic normal flat spaces.)

It may be finally remarked that for any space-time satisfying the weak energy condition [11], every null geodesic γ gives rise to a plane wave W_γ admitting an electromagnetic field tensor which, together with the metric of W_γ constitutes a (null) solution of the Einstein-Maxwell equations (cf. Lichnerowicz [13]). This is because the metric of W_γ is governed by three free real functions of the affine parameter x^0 on γ (whose information is contained in C_{22}, C_{23} and C_{33}). Two of these functions determine the Weyl tensor of W_γ and arise from the complex gravitational null datum Ψ_0 on γ (cf. [14], take l^a tangent to γ); the third function determines the Ricci tensor of W_γ and arises from the Ricci component Φ_{00} on γ (cf. [14]). The weak energy condition gives $\Phi_{00} \geqslant 0$, so we can write $\Phi_{00} = |\varphi_0|^2$ where φ_0 is a complex function of x^0. The quantities Ψ_0 and φ_0 define the strengths of the null gravitational and null electromagnetic field, respectively, for the plane wave W_γ. They can be chosen as arbitrary functions of x^0 to give a solution of the Einstein-Maxwell equations. Note that W_γ is a solution of the vacuum equations for every γ if and only if $\Phi_{00} = 0$ along every γ, i.e. if and only if M is an Einstein space (possibly with non-zero cosmological constant).

104 Any space-time has a plane wave as a limit; *Differential Geometry and Relativity.* Eds M. Cahen and M. Flato, pp. 271–5. *Mathematical Phys. and Appl. Math.*, Vol. 3, Reidel, Dordrecht, 1976

REFERENCES

[1] Penrose, R., *Techniques of Differential Topology in Relativity*, SIAM, Philadelphia, 1972.
[2] Rosen, N., *Phys. Z. Sowjet.* **12** (1937), 366.
[3] Bondi, H., Pirani, F. A. E., and Robinson, I., *Proc. Roy. Soc. (Lond.)* **A251** (1959), 519.
[4] Geroch, R., *Commun. Math. Phys.* **13** (1969), 180.
[5] Penrose, R., *Ann. Phys.* **10** (1960), 171.
[6] Pirani, F. A. E., *Proc. Roy. Soc. (Lond.)* **A252** (1959), 96.
[7] Synge, J. L., *Relativity: The Special Theory*, North-Holland Publ. Co., Amsterdam, 1956.
[8] Robinson, I., Report to the Eddington Group, Cambridge, 1956.
[9] Brinkman, H. W., *Proc. Natl. Acad. Sci. (U.S.)* **9** (1923), 1.
[10] Penrose, R., *Revs. Mod. Phys.* **37** (1965), 215.
[11] Hawking, S. W. and Ellis, G. F. R., *The Large Scale Structure of Space-Time*, Cambridge Univ. Press, 1973.
[12] Hawking, S. W. and Penrose, R., *Proc. Roy. Soc. (Lond.)* **A314** (1970), 529.
[13] Lichnerowicz, A., *Compt. Rend.* **246** (1958), 893.
[14] Newman, E. T. and Penrose, R., *J. Math. Phys.* **3** (1962), 566.

84
THE KÄHLER STRUCTURE OF ASYMPTOTIC TWISTOR SPACE

J. Math. Phys. **18** (1977), no. 1, 58–64. With M. Ko and E. T. Newman

The Kähler structure of asymptotic twistor space

M. Ko and E. T. Newman*

Department of Physics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

R. Penrose

Mathematics Institute, Oxford University, Oxford, England
(Received 2 September 1976)

Asymptotic twistor space $\mathcal{T}$ is a 4-complex-dimensional Kähler manifold (of signature $++--$) which can be constructed from an asymptotically flat space–time containing gravitational radiation. The properties of this Kähler structure are investigated, the Kähler metric being of a particular type, arising from a scalar Σ with special homogeneity properties. The components of the Kähler curvature $K^{\alpha\dot\beta}{}_{\gamma\dot\delta}$ are found explicitly in terms of the asymptotic Weyl curvature of the space–time. When gravitational radiation is present, $K^{\alpha\dot\beta}{}_{\gamma\dot\delta}\neq 0$, whereas for a stationary field $K^{\alpha\dot\beta}{}_{\gamma\dot\delta}=0$. The "Ricci-flat" condition $K^{\alpha\dot\beta}{}_{\alpha\dot\gamma}=0$ is found always to hold.

INTRODUCTION

The theory of twistors provides a novel approach to the treatment of special-relativistic physics which is particularly well suited to the description of massless fields (of any spin) and other conformally invariant aspects of physics.[1,2] The theory can also be used to treat situations in which conformal invariance does not hold and when the space–time is not flat but, as yet, little work has been published on this. In the present paper we examine the curved twistor space $\mathcal{T}$ which arises naturally from an asymptotically flat space–time $\mathcal{M}$ in which gravitational radiation may be present. It is the presence of this radiation which modifies the structure of the twistor space from the standard one arising in the case of flat Minkowski space M. This standard twistor space T is a complex four-dimensional vector space, on which is defined a Hermitian form Σ of signature $(++--)$. [This structure for T is invariant under the fifteen parameter conformal group $C^{\uparrow}_{+}(1,3)$ for M, in accordance with the local isomorphism $SU(2,2)\to C^{\uparrow}_{+}(1,3)$.] But when (outgoing) radiation is present the vector space structure of the twistor space is lost. However, $\mathcal{T}$ remains a complex manifold; furthermore it retains a real scalar field Σ defined analogously to the standard Σ for T, but which cannot now be regarded as a Hermitian form owing to the lack of linear structure for $\mathcal{T}$.[1] The second derivatives of Σ (once holomorphic and once antiholomorphic) with respect to the coordinates in $\mathcal{T}$ define the components of a Kähler metric on $\mathcal{T}$ [of signature $(++--)$]. The curvature $K^{\alpha\dot\beta}_{\gamma\dot\delta}$ of this Kähler metric had been computed earlier, using a combination of algebraic and geometric arguments, the result being given in barest outline in Ref. 1. It turns out that only five of the 100 independent curvature components survive, that $K^{\alpha\dot\beta}_{\gamma\dot\delta}$ is Ricci-flat (i.e., $K^{\alpha\dot\beta}_{\alpha\dot\delta}=0$), and that these surviving components can all be expressed explicitly in terms of standard asymptotic curvature quantities for the space–time $\mathcal{M}$ which describe the radiation behavior of the gravitational field. These earlier arguments, somewhat involved conceptually and difficult to express explicitly, have not been described in print. Here we present a much more direct approach to the problem, which shows how these curvature components may be obtained by a straightforward computation.

The discussion given here, as regards twistor theory, will be essentially self-contained. Some familiarity with the standard notation for asymptotically flat space–times will be assumed.

1. KÄHLER MANIFOLDS

The definition of a Kähler manifold K can be given in various different equivalent forms.[3–5] For our present purposes, K is a complex n-dimensional manifold[6] on which is defined a nondegenerate Hermitian metric $\underline{G}$, obtainable (locally) as a second derivative (once holomorphic and once antiholomorphic) of a real scalar field Σ. Thus, if z^{α} $(\alpha=1,\ldots,n)$ are local holomorphic coordinates for K, with complex conjugates $\bar z^{\alpha'}$ $(\alpha'=1',\ldots,n')$, then the components of $\underline{G}$ in this coordinate system constitute the nonsingular Hermitian matrix[7]

$$\begin{bmatrix} G_{11'} & \cdots & G_{1n'} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ G_{n1'} & \cdots & G_{nn'} \end{bmatrix} = G_{\alpha\beta'} = \frac{\partial^2\Sigma}{\partial z^{\alpha}\partial\bar z^{\beta'}}\,. \tag{1.1}$$

Any vector field $\underline{V}$ (real or complex) can be expressed in these coordinates as

$$\underline{V} = V^{\alpha}\frac{\partial}{\partial z^{\alpha}} + W^{\alpha'}\frac{\partial}{\partial\bar z^{\alpha'}}\,. \tag{1.2}$$

The vector field is *real* if

$$W^{\alpha'} = \overline{V^{\alpha}}\quad (=\overline{V^{\alpha}}). \tag{1.3}$$

It is called *holomorphic* if $W^{\alpha'}=0$ and V^{α} are holomorphic functions of z^{β} (i.e., $\partial V^{\alpha}/\partial\bar z^{\beta'}=0$); it is called *antiholomorphic* if $V^{\alpha}=0$ and $W^{\alpha'}$ are holomorphic in $\bar z^{\beta'}$ (so the complex conjugate of a holomorphic vector field is antiholomorphic). A scalar field ϕ is holomorphic if simply $\partial\phi/\partial\bar z^{\alpha'}=0$ and antiholomorphic if $\partial\phi/\partial z^{\alpha}=0$.

For many purposes it is convenient to use a single index for the entire set of coordinates $z^1,\ldots,z^n,\bar z^{1'},\ldots,\bar z^{n'}$ and write

$$V^{\Gamma} = (V^{\alpha}, W^{\alpha'})\quad (\Gamma=1,\ldots,n,1',\ldots,n'). \tag{1.4}$$

Then the components

$$G_{\Gamma\Delta}=\begin{bmatrix}0 & G_{\gamma\lambda'}\\ G_{\lambda\gamma'} & 0\end{bmatrix} \tag{1.5}$$

define a (real) Riemannian metric on K. This metric is real because

$$G_{\Gamma\Delta}V^{\Gamma}V^{\Delta}=G_{\gamma\lambda'}V^{\gamma}W^{\lambda'}+G_{\lambda\gamma'}W^{\gamma'}V^{\lambda} \tag{1.6}$$

is real whenever $\underline{V}$ is real. A Riemann curvature tensor $\underline{R}$ is thereby defined, with components $R_{\Gamma\Delta\Sigma\Theta}$, the only nonvanishing members turning out to be those of the form

$$R_{\gamma\lambda'\xi\theta'},\quad R_{\gamma\lambda'\theta'\xi},\quad R_{\gamma'\lambda\xi\theta'},\quad R_{\gamma'\lambda\xi'\theta}. \tag{1.7}$$

The Kähler metric components $G_{\alpha\beta'}$, together with their inverse matrix $G^{\alpha\beta'}$ for which

$$G_{\alpha\beta'}G^{\gamma\beta'}=\delta^{\gamma}_{\alpha},\quad G^{\alpha\beta'}G_{\alpha\gamma'}=\delta^{\beta'}_{\gamma'}, \tag{1.8}$$

can be used to eliminate all primed indices from a tensor expression by pushing them all into the opposite position, e.g.,

$$T_{\alpha\beta\gamma'}{}^{\rho\sigma'}=T_{\alpha\beta}{}^{\gamma\rho}{}_{\sigma}=T_{\alpha\beta\mu'}{}^{\rho\nu'}G^{\gamma\mu'}G_{\sigma\nu'}. \tag{1.9}$$

The surviving Riemann tensor components (1.7) can then be expressed in terms of the quantity $K^{\alpha\beta}_{\lambda\mu}$

$$K^{\alpha\beta}_{\lambda\mu}=R_{\lambda}{}^{\alpha}{}_{\mu}{}^{\beta}=-R_{\lambda}{}^{\alpha\beta}{}_{\mu}=-R^{\alpha}{}_{\lambda\mu}{}^{\beta}=R^{\alpha}{}_{\lambda}{}^{\beta}{}_{\mu}. \tag{1.10}$$

It turns out (as we shall see shortly) that $K^{\alpha\beta}_{\lambda\mu}$ enjoys a symmetry and a Hermiticity property:

$$\overline{K^{\alpha\beta}_{\lambda\mu}}=\overline{K^{\alpha'\beta'}_{\lambda'\mu'}}=K^{\alpha\beta}_{\lambda\mu}=K^{(\alpha\beta)}_{(\lambda\mu)}. \tag{1.11}$$

Calculations can be made independent of coordinates by the introduction of an abstract index formulation. Indices $\alpha, \beta, \ldots, \omega, \alpha_0, \ldots, \alpha_1, \cdots$ are used here abstractly (so V^{α} *is* a vector, namely the vector with components V^{α}—the "primed" components being taken as zero). The corresponding *conjugate* abstract indices $\alpha', \beta', \cdots$ are *not* needed since we can write W_{α} for the vector with components $W^{\beta'}G_{\alpha\beta'}$. Thus, instead of giving rise to primed abstract indices, compex conjugation entails the interchange of all upper and lower indices, e.g.,

$$\overline{T^{\alpha\beta}}_{\gamma}=\overline{T}_{\alpha\beta}{}^{\gamma}, \tag{1.12}$$

which expresses the fact that $\overline{T^{\alpha\beta}}_{\gamma}=\overline{T}^{\alpha'\beta'}{}_{\gamma'}$ corresponds to $\overline{T}_{\alpha\beta}{}^{\gamma}=\overline{T}^{\alpha'\beta'}{}_{\gamma'}G_{\alpha\alpha'}G_{\beta\beta'}G^{\gamma\gamma'}$. The abstract version of the Kähler metric tensor is, with this notation, simply the Kronecker tensor δ^{β}_{α}, as follows from (1.8). The Kähler Hermitian scalar product between vectors U^{α}, V^{α} is thus

$$U^{\alpha}\overline{V}_{\alpha}=G_{\alpha\beta'}U^{\alpha}\overline{V}^{\beta'}. \tag{1.13}$$

The notation is thereby brought into line with that of standard twistor theory.[1,2]

The information concerning the curvature, etc., of the Kähler structure goes into the formal properties of covariant derivative. In components we have

$$V^{\Theta}{}_{;\Phi}=V^{\Theta}{}_{,\Phi}+\Gamma^{\Theta}_{\Psi\Phi}V^{\Psi}, \tag{1.14}$$

where the $\Gamma^{\Theta}_{\Psi\Phi}$ are the ordinary Christoffel symbols for the metric (1.5). From (1.1) and (1.5) one obtains

$$\Gamma^{\phi}_{\phi\phi}=G^{\phi\chi'}\frac{\partial^3\Sigma}{\partial z^{\phi}\partial z^{\phi}\partial\bar{z}^{\chi'}},\quad \Gamma^{\phi'}_{\phi'\phi'}=G^{\chi\phi'}\frac{\partial^3\Sigma}{\partial z^{\chi}\partial\bar{z}^{\phi'}\partial\bar{z}^{\phi'}},$$

$$\Gamma^{\phi}_{\phi\phi'}=\Gamma^{\phi}_{\phi'\phi}=\Gamma^{\phi'}_{\phi\phi}=\Gamma^{\phi'}_{\phi'\phi'}=\Gamma^{\phi'}_{\phi'\phi}=\Gamma^{\phi'}_{\phi\phi'}=0 \tag{1.15}$$

so that

$$V^{\phi}{}_{;\phi}=\frac{\partial V^{\phi}}{\partial z^{\phi}}+\Gamma^{\phi}_{\phi\phi}V^{\phi},\quad V^{\phi;\phi}=\frac{\partial V^{\phi}}{\partial\bar{z}^{\phi'}}G^{\phi\phi'},$$

$$W_{\theta;\phi}=\frac{\partial W_{\theta}}{\partial z^{\phi}}-\Gamma^{\phi}_{\theta\phi}W_{\phi},\quad W_{\theta}{}^{;\phi}=\frac{\partial W_{\theta}}{\partial\bar{z}^{\phi'}}G^{\phi\phi'}. \tag{1.16}$$

The four expressions (1.16) are the component forms of the respective abstract expressions[9]

$$\Box_{\phi}V^{\phi},\quad \Box^{\phi}V^{\phi},\quad \Box_{\phi}W_{\theta},\quad \Box^{\phi}W_{\theta}. \tag{1.17}$$

The two derivative operators $\Box_{\phi}$, $\Box^{\phi}$ extend so as to apply to any tensor quantity $T^{\lambda\ldots\nu}_{\alpha\ldots\gamma}$ in the usual way, so that the normal additivity and Leibniz properties hold. The relation to the curvature $K^{\alpha\beta}_{\lambda\mu}$ is achieved via

$$(\Box_{\lambda}\Box^{\alpha}-\Box^{\alpha}\Box_{\lambda})V^{\beta}=K^{\alpha\beta}_{\lambda\mu}V^{\mu},$$

$$(\Box^{\lambda}\Box_{\alpha}-\Box_{\alpha}\Box^{\lambda})W_{\beta}=K^{\lambda\mu}_{\alpha\beta}W_{\mu}, \tag{1.18}$$

as follows from (1.10) and the relation

$$V^{\Theta}{}_{;\Phi;\Psi}-V^{\Theta}{}_{;\Psi;\Phi}=R^{\Theta}{}_{\Sigma\Phi\Psi}V^{\Sigma}. \tag{1.19}$$

Alternatively, the coordinate expressions

$$K_{\alpha\beta'\lambda\mu'}=-G_{\lambda\mu'},_{\alpha\beta'}+G^{\rho\sigma'}G_{\rho\mu'},_{\beta'}G_{\lambda\sigma'},_{\alpha}, \tag{1.20}$$

$$K_{\alpha\beta'\lambda}{}^{\mu}=-\Gamma^{\mu}_{\alpha\lambda,\beta'} \tag{1.21}$$

may be employed.

We have the vanishing torsion properties

$$\Box_{\alpha}\Box^{\beta}f=\Box^{\beta}\Box_{\alpha}f \tag{1.22}$$

and

$$\Box_{[\alpha}\Box_{\beta]}f=0,\quad \Box^{[\alpha}\Box^{\beta]}f=0 \tag{1.23}$$

for any scalar f. In fact, relations (1.23) generalize to the commutation properties

$$\Box_{\alpha}\Box_{\beta}=\Box_{\beta}\Box_{\alpha},\quad \Box^{\alpha}\Box^{\beta}=\Box^{\beta}\Box^{\alpha} \tag{1.24}$$

applied to *any* tensor quantity. Equation (1.24) follows because (1.7) are the only surviving Riemann tensor components. But a simple "abstract" proof of these facts can also be given. Choose f to be an arbitrary holomorphic scalar on K. Then $\Box^{\alpha}f=0$ (since the component version of $\Box^{\alpha}f$ is $G^{\alpha\beta'}\partial f/\partial\bar{z}^{\beta'}$), whereas $\Box_{\alpha}f$ is arbitrary at any one point. Since the torsion vanishes, we would have a curvature quantity $Q^{\alpha\beta\gamma}_{\lambda}$ for which

$$(\Box^{\alpha}\Box^{\beta}-\Box^{\beta}\Box^{\alpha})\Box_{\lambda}f=Q^{\alpha\beta\gamma}_{\lambda}\Box_{\gamma}f. \tag{1.25}$$

But

$$\Box^{\alpha}\Box^{\beta}\Box_{\lambda}f=\Box^{\alpha}\Box_{\lambda}\Box^{\beta}f=0,$$

and similarly if α and β are interchanged. Thus (1.25) vanishes, and so $Q^{\alpha\beta\gamma}_{\lambda}=0$ because of the arbitrariness at each point of $\Box_{\gamma}f$. The second of relations (1.24) follows—and the proof of the first is similar.

The symmetry properties of $K^{\alpha\beta}_{\lambda\mu}$ also follow directly from the formal properties of $\sqcup_\alpha$ and $\sqcap^\alpha$. The "Hermiticity"

$$\bar{K}^{\alpha\beta}_{\lambda\mu} = K^{\alpha\beta}_{\lambda\mu} \tag{1.26}$$

[cf. (1.11)] is a direct consequence of (1.18) and the fact that $\sqcup_\alpha$ and $\sqcap^\alpha$ are complex conjugates of one another. (In components, $\overline{\sqcup_\alpha} = G_{\beta\alpha'}\sqcap^\beta$.) The symmetry of $K^{\lambda\mu}_{\alpha\beta}$ in α, β follows because

$$\begin{aligned} K^{\lambda\mu}_{\alpha\beta}\sqcup_\mu f &= (\sqcap^\lambda\sqcup_\alpha - \sqcup_\alpha\sqcap^\lambda)\sqcup_\beta f \\ &= \sqcap^\lambda\sqcup_\alpha\sqcup_\beta f - \sqcup_\alpha\sqcup_\beta\sqcap^\lambda f, \end{aligned}$$

each term of which is symmetric in α and β, by (1.24) (whether or not f is chosen holomorphic). The symmetry in λ, μ follows similarly, so that

$$K^{\lambda\mu}_{\alpha\beta} = K^{(\lambda\mu)}_{(\alpha\beta)}. \tag{1.27}$$

The "Bianchi identities" are correspondingly derived:

$$\begin{aligned} \sqcup_{[\alpha}(K^{\lambda}_{\beta]\gamma}{}^{\mu}W_\mu) &= \sqcup_{[\alpha}(\sqcap^\lambda\sqcup_{\beta]} - \sqcup_{\beta]}\sqcap^\lambda)W_\gamma \\ &= \sqcap^\lambda\sqcup_{[\alpha}\sqcup_{\beta]}W_\gamma - K^{\lambda\mu}_{[\alpha\beta]}\sqcup_\mu W_\gamma - K^{\lambda\mu}_{\gamma[\alpha}\sqcup_{\beta]}W_\mu - 0 \end{aligned}$$

gives, on expanding the left-hand side and using (1.24), (1.27),

$$\sqcup_{[\alpha}K^{\lambda}_{\beta]\gamma}{}^{\mu} = 0 \tag{1.28}$$

and, similarly

$$\sqcap^{[\lambda}K^{\mu]\nu}_{\alpha\beta} = 0. \tag{1.29}$$

Note that by (1.27) we can re-express these relations as

$$\sqcup_\alpha K^{\lambda\mu}_{\beta\gamma} = \sqcup_{(\alpha}K^{\lambda\mu}_{\beta\gamma)}, \quad \sqcap^\lambda K^{\mu\nu}_{\alpha\beta} = \sqcap^{(\lambda}K^{\mu\nu)}_{\alpha\beta}. \tag{1.30}$$

2. KÄHLER MANIFOLDS WITH HOMOGENEOUS SCALAR

All the above properties hold in a general Kähler manifold. The scalar Σ need not be globally defined and is not, itself, considered to be part of the Kähler structure of $\mathcal{K}$. However, in the situation that concerns us here, Σ *has geometric meaning*. Furthermore, it has a certain homogeneity property which can be described as follows: There exist coordinates z^α, $\bar{z}^{\alpha'}$ (possibly defined only locally) for which

$$\Sigma = \Sigma(z^\alpha, \bar{z}^{\alpha'}) \tag{2.1}$$

is separately homogeneous of degree unity in z^α and in $\bar{z}^{\alpha'}$, i.e.,

$$\Sigma(\lambda z^\alpha, \bar{\lambda}\bar{z}^{\alpha'}) = \lambda\bar{\lambda}\Sigma(z^\alpha, \bar{z}^{\alpha'}). \tag{2.2}$$

Euler's theorem gives

$$z^\alpha\Sigma_{,\alpha} = \Sigma = \bar{z}^{\alpha'}\Sigma_{,\alpha'} \tag{2.3}$$

and, by taking one more derivative [cf. (1.1)],

$$z^\alpha G_{\alpha\beta'} = \Sigma_{,\beta'}, \quad \bar{z}^{\alpha'}G_{\beta\alpha'} = \Sigma_{,\beta}. \tag{2.4}$$

We introduce vectors Z^α and $\bar{Z}_\alpha$, whose components are z^α and $\bar{z}^{\beta'}G_{\alpha\beta'}$, respectively, whence by (2.4) we have

$$Z^\alpha = \sqcap^\alpha\Sigma, \quad \bar{Z}_\alpha = \sqcup_\alpha\Sigma. \tag{2.5}$$

Equation (2.3) gives

$$\Sigma = Z^\alpha\bar{Z}_\alpha. \tag{2.6}$$

A tensor $T^{\alpha\cdots\tau}_{\alpha\cdots\gamma}$ will be called *homogeneous of degree* (p, q) if its components satisfy

$$T^{\alpha\cdots\tau}_{\alpha\cdots\gamma}(\lambda z^\theta, \bar{\lambda}\bar{z}^{\theta'}) = \lambda^p\bar{\lambda}^q T^{\alpha\cdots\tau}_{\alpha\cdots\gamma}(z^\theta, \bar{z}^{\theta'}). \tag{2.7}$$

By referring to (1.16) and using the facts

$$\begin{aligned} &z^\Phi\Gamma^\Theta_{\Phi\Phi} = (G^{\Theta X'}\Sigma_{,\Phi X'})_{,\Phi}z^\Phi = 0, \\ &G^{\Phi\Phi'}{}_{,X}z^X = 0 \end{aligned} \tag{2.8}$$

(which follow, again, from Euler's theorem) we have the "covariant" version of Euler's theorem

$$\begin{aligned} Z^\mu\sqcup_\mu T^{\alpha\cdots\tau}_{\alpha\cdots\gamma} &= pT^{\alpha\cdots\tau}_{\alpha\cdots\gamma}, \\ \bar{Z}_\mu\sqcap^\mu T^{\alpha\cdots\tau}_{\alpha\cdots\gamma} &= qT^{\alpha\cdots\tau}_{\alpha\cdots\gamma}, \end{aligned} \tag{2.9}$$

expressing the homogeneity (2.7). Since, by (2.5), Z^α and $\bar{Z}_\alpha$ are defined explicitly from Σ, $\sqcup_\alpha$, and $\sqcap^\alpha$, we see that (2.9) expresses the homogeneity (2.7) in a coordinate independent fashion. One readily obtains the homogeneity degrees

$$\begin{aligned} &\Sigma\colon (1,1), \quad Z^\alpha\colon (1,0), \quad \bar{Z}_\alpha\colon (0,1), \\ &\sqcup_\alpha\colon (-1,0), \quad \sqcap^\alpha\colon (0,-1), \quad K^{\alpha\beta}_{\rho\sigma}\colon (-1,-1), \end{aligned} \tag{2.10}$$

where the homogeneity of the operators $\sqcup_\alpha$, $\sqcap^\alpha$ is to be interpreted as the fact that when acting on a tensor of homogeneity degrees (p, q), $\sqcup_\alpha$ produces one of degrees $(p-1, q)$ and $\sqcap^\alpha$ produces one of degrees $(p, q-1)$.

Since $Z^\alpha{}_{;\beta} = z^\alpha{}_{,\beta} = \delta^\alpha_\beta$ [by (1.16) and (2.8)] and $Z^\alpha{}_{;\beta'} = z^\alpha{}_{,\beta'} = 0$ (together with their conjugates), we have

$$\sqcup_\beta Z^\alpha = \delta^\alpha_\beta = \sqcap^\alpha\bar{Z}_\beta \tag{2.11}$$

and

$$\sqcap^\alpha Z^\beta = 0, \quad \sqcup_\alpha\bar{Z}_\beta = 0. \tag{2.12}$$

Equations (2.11), (2.12) tell us that $(Z^\alpha, \bar{Z}_\alpha)$ behaves like a "position vector" for the space $\mathcal{K}$. Indeed, writing $z^\Theta = (z^\theta, \bar{z}^{\theta'})$, we can re-express (2.11), (2.12) as

$$z^\Theta{}_{;\Phi} = \delta^\Theta_\Phi, \tag{2.13}$$

the semicolon now denoting the full operation (1.14).

This analogy with the concept of position vector is brought out more strongly if we adopt a somewhat different attitude to the space $\mathcal{K}$ which is useful in some contexts. According to this alternative view we allow the coordinates z^α and $\bar{z}^{\alpha'}$ to vary independently of one another—so instead of using the notation $\bar{z}^{\alpha'}$ for the second set of coordinates, we introduce new independent quantities[10] $w^{\alpha'}$, where $w^{\alpha'}$ ranges over some suitably small neighborhood of $\bar{z}^{\alpha'}$, for each z^α. The manifold now becomes $2n$-complex-dimensional (i.e., $4n$-real-dimensional) with local holomorphic coordinates z^α, $w^{\alpha'}$. We denote this new space by $\mathbb{C}\mathcal{K}$; it is a complexification ("second complexification"!) of $\mathcal{K}$ considered as a *real* $2n$-dimensional manifold. The "real slice" $\mathcal{K}$ of $\mathbb{C}\mathcal{K}$ is given when $w^{\alpha'} = \bar{z}^{\alpha'}$. We suppose that Σ is real-analytic on $\mathcal{K}$, i.e., analytic in the real and imaginary parts of z^α so that its definition extends *holomorphically* (i.e., complex-analytically) to $\mathbb{C}\mathcal{K}$, which is assumed to be a sufficiently small complex extension of $\mathcal{K}$ that this is uniquely possible. Thus,

$$\Sigma = \Sigma(z^\alpha, w^{\alpha'}) \tag{2.14}$$

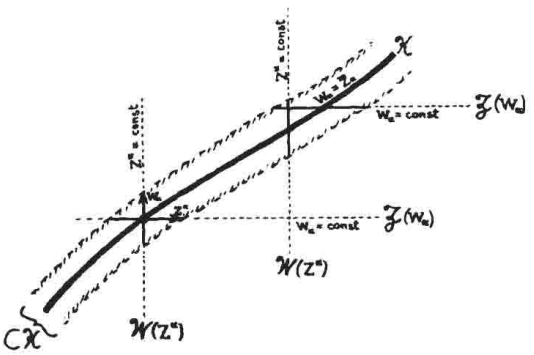


FIG. 1.

is a holomorphic function (for each coordinate chart) of $2n$ complex variables.

The Kähler metric and the covariant derivative operators $\square_\alpha$, $\square^\alpha$ also extend uniquely to $C\mathcal{K}$ with the same formal properties as before, but now we write W_α instead of $\bar{Z}_\alpha$:

$$Z^\alpha = \square^\alpha \Sigma, \quad W_\alpha = \square_\alpha \Sigma, \quad Z^\alpha W_\alpha = \Sigma, \tag{2.15}$$

$$\square_\alpha Z^\beta = \delta_\alpha^\beta = \square^\beta W_\alpha, \quad \square^\alpha Z^\beta = 0, \quad \square_\alpha W_\beta = 0. \tag{2.16}$$

We may think of

$$Z^\alpha \oplus W_\alpha \tag{2.17}$$

as denoting the "position vector" of a point of $C\mathcal{K}$. As we hold W_α fixed and vary Z^α, we obtain a space which has a well-defined structure as a vector-space $\mathcal{L}(W_\alpha)$. The existence of this vector space structure follows because the operators $\square_\alpha$ *commute* with each other (1.24) and satisfy the first relation (2.16). The vector Z^α is now legitimately a position vector (in the normal sense) for the space $\mathcal{L}(W_\alpha)$ for each given W_α. Similarly, as we hold Z^α fixed and vary W_α, we obtain a vector space $\mathcal{W}(Z^\alpha)$ for which W_α is a normal position vector and $\square^\alpha$ the corresponding gradient operator (see Fig. 1). The space $C\mathcal{K}$ itself does not, however, have a vector space structure, in general. Although (2.17) behaves formally like a position vector, with respect to the operator

$$\square_\alpha \oplus \square^\alpha, \tag{2.18}$$

[cf. (2.13)], the operators (2.18) do not commute with one another, the commutation giving rise to the curvature $K_{\rho\sigma}^{\alpha\beta}$. Thus, $K_{\rho\sigma}^{\alpha\beta}$ expresses the way in which the flat structure of each $\mathcal{L}(W_\alpha)$ gets "twisted around" in relation to its neighbors, i.e., as W_α varies, and the way in which the flat structure of each $\mathcal{W}(Z^\alpha)$ gets "twisted around" as Z^α varies.

As a final remark in this section, we note

$$(\square_\rho \square^\alpha - \square^\alpha \square_\rho) Z^\beta = 0 - \square^\alpha \delta_\rho^\beta = 0$$

so that

$$K_{\rho\sigma}^{\alpha\beta} Z^\sigma = 0$$

and similarly

$$K_{\rho\sigma}^{\alpha\beta} W_\beta = 0,$$

which takes the form

$$K_{\rho\sigma}^{\alpha\beta} \bar{Z}_\beta = 0$$

on $\mathcal{K}$.

3. ASYMPTOTIC TWISTOR SPACE

We now specialize the preceding theory to the case of the space $\mathcal{T}$ of asymptotic twistors. This space has been introduced and briefly discussed in Refs. 1 and 2. Here we give an independent derivation of the properties of $\mathcal{T}$.

Let $\mathcal{M}$ be an asymptotically flat space–time, with future[11] null conformal infinity $\mathcal{I}^+$. Suppose that $\mathcal{M}$ and its metric g_{ab} are such that a conformally related metric $\hat{g}_{ab} = \Omega^2 g_{ab}$ exists for $\bar{\mathcal{M}}$ $(= \mathcal{M} \cup \mathcal{I}^+)$ which is analytic in a neighborhood of every point of $\mathcal{I}^+$, the conformal factor Ω being also analytic. Let $C\bar{\mathcal{M}}$ and $C\mathcal{I}^+$ denote complexifications of $\bar{\mathcal{M}}$ and $\mathcal{I}^+$, respectively. (At the present state of understanding it is unfortunately necessary to assume analyticity for the space–time $\mathcal{M}$. It is to be hoped that future developments will allow this restriction to be removed. In any case, a nonanalytic space–time can be approximated arbitrarily closely by analytic ones. It is not necessary that the complexification $C\bar{\mathcal{M}}$ extend very "far" from the real section $\bar{\mathcal{M}}$; the radius of convergence of the Taylor series at each point P of $\bar{\mathcal{M}}$ need only be some positive quantity $\epsilon(P)$, and then $\epsilon(P)$ will also define how far into the complex $C\bar{\mathcal{M}}$ extends at P.)

Certain cross-sections of $C\mathcal{I}^+$ have been referred to[12,13] as "good cuts". These are the intersections of $C\mathcal{I}^+$ with null hypersurfaces $\mathcal{N}$ in $C\bar{\mathcal{M}}$ whose asymptotic *shear*—taking the appropriate measure of shear—vanishes. There are, in fact, two measures of asymptotic shear for $\mathcal{N}$, denoted σ^0 and $\tilde{\sigma}^0$. For a real null hypersurface $\mathcal{N}$, we should have had $\tilde{\sigma}^0 = \bar{\sigma}^0$, but for a complex $\mathcal{N}$, σ^0 and $\tilde{\sigma}^0$ are independent quantities. We have the choice to make either $\sigma^0 = 0$ or $\tilde{\sigma}^0 = 0$ (but not both, if outgoing gravitational radiation is present). The choice $\sigma^0 = 0$ has been normally made in earlier work. Here we select $\tilde{\sigma}^0$, instead, since this fits in rather better with the notation for twistor space.

Let u, ζ, and $\tilde{\zeta}$ be complex *Bondi-type coordinates* (cf. Refs. 12, 13) for $C\mathcal{I}^+$, the points of $\mathcal{I}^+$ being given by $\tilde{\zeta} = \bar{\zeta}$, u real ($\zeta = \infty$ being allowed). A "good" cross section $\mathcal{X}$ of $C\mathcal{I}^+$ is defined by a "cut function" $X = X(\zeta, \tilde{\zeta})$, satisfying the equation

$$\tilde{\delta}^2 X = \tilde{\sigma}^0(X, \zeta, \tilde{\zeta}), \tag{3.1}$$

where $\mathcal{X}$ is defined by

$$u = X(\zeta, \tilde{\zeta}). \tag{3.2}$$

The quantity ζ, in Eq. (3.1), appears only as a parameter; for each fixed value ζ_0 of ζ, (3.1) defines a two-complex-parameter family of curves in $(u, \tilde{\zeta})$-space which are, in fact, null geodesics on $C\mathcal{I}^+$. As ζ_0 varies, we therefore get a three-complex-parameter family. Each such null geodesic γ defines an *asymptotic twistor* Z^α, but only *up to proportionality*.

For the *twistor* Z^α proper (as opposed to the projective twistor) we require also a "scaling" for γ. This scaling will be necessary in order to assign a meaning to $\Sigma = Z^\alpha \bar{Z}_\alpha$. We have one more complex parameter for

the scaling, so the space $\mathcal{T}$ of asymptotic twistors Z^α will be four-complex-dimensional. This scaling is actually achieved by assigning a spinor $\pi_{A'}$ at each point of γ, which is parallelly propagated along γ, and for which $\iota^A\pi^{A'}$ is a tangent vector to γ. The tangent vectors to the generators of $C\mathcal{I}^+$ (corresponding to the parameter u) are n^a where $n^a=\iota^A\tilde{\iota}^{A'}$. For any standard choice of ι^A and corresponding $\tilde{\iota}^{A'}$ (where we normally take ι^A and $\tilde{\iota}^{A'}$ parallelly propagated along generators of $C^+\mathcal{I}$, with $\tilde{\iota}^{A'}=\bar{\iota}^{A'}$ on $\mathcal{I}^+$), the complex number

$$\pi_{1'}=\pi_{A'}\tilde{\iota}^{A'} \tag{3.3}$$

can be used to measure the required scaling for Z^α.

To obtain the definition of Σ, we need to locate $\bar{\gamma}$, the complex conjugate geodesic to γ on $C\mathcal{I}^+$. Now any locus on $C\mathcal{I}^+$ will have a uniquely defined complex conjugate locus, namely that which is obtained from the original locus when the point transformation of complex conjugation is applied to $C\mathcal{I}^+$. If real coordinates had been chosen for $\mathcal{I}^+$, this would simply be the transformation sending each point of $C\mathcal{I}^+$ into the one with the complex conjugated coordinates. But with the present choice, the complex conjugate of the point labelled $(u, \zeta, \tilde{\zeta})$ is labelled $(\bar{u}, \tilde{\zeta}, \bar{\zeta})$. Thus, $\bar{\gamma}$ lies on the (u, ζ)-plane defined by $\tilde{\zeta}=\zeta_0$. Furthermore, assuming that γ is such that u attains a well-defined value u_0 on it when $\tilde{\zeta}=\bar{\zeta}_0$, then γ and $\bar{\gamma}$ will each meet a (unique) common generator μ of $C\mathcal{I}^+$, namely that defined by $\zeta=\zeta_0$, $\tilde{\zeta}=\bar{\zeta}_0$. The two points of intersection of γ and $\bar{\gamma}$ with μ will then have respective u values u_0 and $\bar{u}_0$. [We note also that $\bar{\gamma}$ is a member of the two complex-parameter family of curves in the (u, ζ)-plane which are defined by solutions $X(\tilde{\zeta}, \zeta)$, at $\tilde{\zeta}=\bar{\zeta}_0$, of $\delta^2X(\tilde{\zeta}, \zeta)=\sigma^0(X, \tilde{\zeta}, \zeta)$.] We now define[1]

$$\Sigma=i(u_0-\bar{u}_0)\,n^{AA'}\bar{\pi}_A\pi_{A'}, \tag{3.4}$$

which, by (3.3) and $n^a=\iota^A\bar{\iota}^{A'}$, can be written

$$\Sigma=i(u_0-\bar{u}_0)\,\bar{\pi}_1\pi_{1'}. \tag{3.5}$$

We shall require a holomorphic parametrization for the system of Z^α's, so that we can take derivatives of Σ with respect to these parameters. The different solutions of (3.1) for fixed $\zeta=\zeta_0$ can be parametrized by pairs of complex numbers μ^0, μ^1 [constant of integration for the second-order ordinary differential equation (3.1)] and we assume that as ζ_0 varies, the solutions (i.e., the curves γ) vary holomorphically as μ^0, μ^1 vary holomorphically. Thus, we have a single holomorphic function

$$X=X(\tilde{\zeta}; \mu^0, \mu^1, \zeta_0) \tag{3.6}$$

of four variables satisfying (3.1), the different solutions of (3.1) being functions of $\tilde{\zeta}$ parametrized by μ^0, μ^1, and ζ_0. Thus μ^0, μ^1, and ζ_0 are holomorphic coordinates for *projective* asymptotic twistor space $P\mathcal{T}$. To assign coordinates to the nonprojective space $\mathcal{T}$, we need to make sure that the scaling for Z^α is also appropriately parametrized. The quantity $\pi_{1'}$ in (3.5) cannot be used as it stands, since it is actually not a holomorphic coordinate, being dependent on the value of $\tilde{\iota}^{A'}$ at the interaction of γ with μ (which depends on $\bar{\gamma}$). However, the spinor $\pi_{A'}$ itself depends holomorphically on Z^α. What is needed is to choose a *component* of $\pi_{A'}$ in a holomorphic way (i.e., in a way "independent" of $\bar{Z}_\alpha$).

To achieve this, we may conformally rescale the metric on $C\mathcal{I}^+$ so that it becomes flat. The original metric on $C\mathcal{I}^+$ is that induced by $\hat{g}_{ab}$ and may be taken to be

$$d\hat{s}=-\frac{4d\zeta\,d\tilde{\zeta}}{(1+\zeta\tilde{\zeta})^2}+0\cdot du^2 \tag{3.7}$$

(cf. Refs. 8, 12). We rescale this to flatness by introducing $d\hat{\hat{s}}^2=(1+\zeta\tilde{\zeta})^2\,d\hat{s}^2$, which is induced on $C\mathcal{I}^+$ by the 4-metric

$$\hat{\hat{g}}_{ab}=(1+\zeta\tilde{\zeta})^2\,\hat{g}_{ab}. \tag{3.8}$$

The replacement of the vector n^a by

$$k^a=(1+\zeta\tilde{\zeta})^{-1}n^a=(1+\zeta\tilde{\zeta})^{-1}\iota^A\tilde{\iota}^{A'}=\kappa^A\tilde{\kappa}^{A'} \tag{3.9}$$

naturally accompanies this rescaling (cf. Ref. 12), and k^a becomes covariantly constant on $C\mathcal{I}^+$ with respect to $\hat{\hat{g}}_{ab}$. We can arrange that κ^A and $\tilde{\kappa}^{A'}$ are also covariantly constant on $C\mathcal{I}^+$ with respect to $\hat{\hat{g}}_{ab}$ (cf. Ref. 14). Now $\pi_{A'}$ is constant along γ with respect to $\hat{\hat{g}}_{ab}$ as well as with respect to $\hat{g}_{ab}$ (see Ref. 1, 2), so the scalar product

$$r=\pi_{A'}\tilde{\kappa}^{A'} \tag{3.10}$$

is constant along γ and does not now depend on the intersection of γ with $\tilde{\zeta}=$ constant plane. In fact, r is a holomorphic coordinate defining the scaling of Z^α (r being now "independent" of $\bar{\gamma}$ or of $\bar{Z}_\alpha$); substituting (3.10) and (3.9) into (3.5), we get (dropping the subscripts on u_0, $\bar{u}_0$, ζ_0 and $\bar{\zeta}_0$)

$$\Sigma=\Sigma(Z^\alpha, \bar{Z}_\alpha)=ir\bar{r}(1+\zeta\tilde{\zeta})(u-\bar{u}) \tag{3.11}$$

for the Kähler scalar on $\mathcal{T}$, or

$$\Sigma=\Sigma(Z^\alpha, W_\alpha)=ir\tilde{r}(1+\zeta\tilde{\zeta})(u-\tilde{u}) \tag{3.12}$$

for the complexified version $C\mathcal{T}$ of this (with $W_\alpha=\bar{Z}_\alpha$), where

$$u=X(\tilde{\zeta}; \mu^0, \mu^1, \zeta), \tag{3.13a}$$

$$\tilde{u}=\tilde{X}(\zeta; \tilde{\mu}^{0'}, \tilde{\mu}^{1'}, \tilde{\zeta}). \tag{3.13b}$$

In the case of Minkowski space M we can choose

$$X(\tilde{\zeta}; \mu^0, \mu^1, \zeta)=(1+\zeta\tilde{\zeta})^{-1}(\mu^0+\mu^1\tilde{\zeta}), \tag{3.14a}$$

$$\tilde{X}(\zeta; \tilde{\mu}^{0'}, \tilde{\mu}^{1'}, \tilde{\zeta})=(1+\tilde{\zeta}\zeta)^{-1}(\tilde{\mu}^{0'}+\tilde{\mu}^{1'}\zeta) \tag{3.14b}$$

so that

$$\Sigma(Z^\alpha, W_\alpha)=ir\tilde{r}(\mu^0+\mu^1\tilde{\zeta}-\tilde{\mu}^{0'}-\tilde{\mu}^{1'}\zeta). \tag{3.15}$$

Then, introducing new coordinates

$$z^0=ir\mu^0, \quad z^1=-ir\mu^1, \quad z^2=r, \quad z^3=-r\zeta \tag{3.16a}$$

$$w^{0'}=-i\tilde{r}\tilde{\mu}^{0'}, \quad w^{1'}=i\tilde{r}\tilde{\mu}^{1'}, \quad w^{2'}=\tilde{r}, \quad w^{3'}=-\tilde{r}\tilde{\zeta}, \tag{3.16b}$$

we obtain the standard flat-space twistor form[1] of the scalar product

$$\Sigma(Z^\alpha, W_\alpha)=z^0w^{2'}+z^1w^{3'}+z^2w^{0'}+z^3w^{1'}. \tag{3.17}$$

The coordinates z^α, as given by (3.16a), may also be used in the general case for a curved twistor space $\mathcal{T}$, in place of r, μ^0, μ^1, ζ. Notice that μ^0, μ^1, and ζ are defined by the ratios of z^0, z^1, z^2, z^3 while r scales all four proportionally. A similar remark holds for $w^{0'}$, $w^{1'}$, $w^{2'}$, $w^{3'}$ in relation to the tilde variables.

Hence, Σ is indeed homogeneous of degrees (1,1) in the two sets of variables as is required for (2.2). The results of Sec. 2 may therefore be applied.

The signature of (3.17), given when $w^{\alpha'} = \bar{z}^{\alpha'}$ is $++--$. So the signature of $G_{\alpha\bar{\beta}'}$ is $++--$ in the case of flat twistor space. The signature remains constant under continuous deformations for which the Kähler metric does not degenerate. The signature of the Kähler metric of $\mathcal{T}$ must therefore also be $++--$.

It is possible to obtain the Kähler curvature of $\mathcal{T}$ in a completely coordinate-free manner, but the details of this argument are hard to express. Consequently we summarize here only the direct coordinate calculation of the components of the curvature tensor. It is convenient to do the calculation in a new holomorphic coordinate system defined on $\mathcal{T}$ as follows:

$$x^0 = \mu^0, \quad x^1 = \mu^1, \quad x^2 = r, \quad x^3 = \zeta. \tag{3.18}$$

(Note that Σ is not homogeneous in these coordinates.) From (3.11) and (1.5), we obtain the components of the metric tensor

$$G_{\Gamma\bar{\Lambda}} = \begin{array}{c} \\ x^0 \\ x^1 \\ x^2 \\ x^3 \\ \bar{x}^{0'} \\ \bar{x}^{1'} \\ \bar{x}^{2'} \\ \bar{x}^{3'} \end{array} \begin{array}{c} \begin{array}{cccc} x^0 & x^1 & x^2 & x^3 \end{array} \\ \left(\begin{array}{cc|cc} & & 0 & B \\ & 0 & & \\ & & \bar{B}^T & A \\ \hline 0 & \bar{B} & & \\ & & 0 & \\ B^T & A^T & & \end{array}\right) \end{array} \qquad \begin{array}{cccc} \bar{x}^{0'} & \bar{x}^{1'} & \bar{x}^{2'} & \bar{x}^{3'} \end{array} \tag{3.19}$$

with

$$A = \begin{pmatrix} i(1+\zeta\bar{\zeta})(X-\bar{X}) & i\bar{r}[(1+\zeta\bar{\zeta})(X-\bar{X})]_{,\bar{\zeta}} \\ ir[(1+\zeta\bar{\zeta})(X-\bar{X})]_{,\zeta} & ir\bar{r}[(1+\zeta\bar{\zeta})(X-\bar{X})]_{,\zeta\bar{\zeta}} \end{pmatrix}, \tag{3.20}$$

$$B = \begin{pmatrix} ir(1+\zeta\bar{\zeta})X_{,\mu^0} & i\bar{r}[(1+\zeta\bar{\zeta})X_{,\mu^0}]_{,\bar{\zeta}} \\ ir(1+\zeta\bar{\zeta})X_{,\mu^1} & i\bar{r}[(1+\zeta\bar{\zeta})X_{,\mu^1}]_{,\bar{\zeta}} \end{pmatrix},$$

where ^{T} denotes matrix transposition, $^-$ denotes complex conjugation, and commas denote partial derivatives as before.

Using the facts that $\tilde{\delta}^2 X(\bar{\xi}; \mu^0, \mu^1, \zeta) = \tilde{\sigma}^0(X, \zeta, \bar{\xi})$ and $\delta^2 \bar{X}(\xi; \bar{\mu}^{0'}, \bar{\mu}^{1'}, \bar{\zeta}) = \sigma^0(\bar{X}, \bar{\zeta}, \xi)$, the components of the curvature tensor given by (1.20) can be written in the form

$$K_{A\bar{3}', B\bar{3}'} = \frac{-ir\bar{r}\ddot{\bar{\sigma}}^0}{1+\zeta\bar{\zeta}} \frac{\partial X}{\partial x^A} \frac{\partial X}{\partial x^B},$$

$$K_{A\bar{3}', 3\bar{3}'} = \frac{-ir\bar{r}}{(1+\zeta\bar{\zeta})^2} (\ddot{\bar{\sigma}}^0 \delta X + \delta\dot{\bar{\sigma}}^0) \frac{\partial X}{\partial x^A},$$

$$K_{3\bar{3}', 3\bar{3}'} = \frac{-ir\bar{r}}{(1+\zeta\bar{\zeta})^3} \times \{\ddot{\bar{\sigma}}^0 (\delta X)^2 + 2\delta X \delta\dot{\bar{\sigma}}^0 + \delta^2 \tilde{\sigma}^0 + \sigma^0 \dot{\tilde{\sigma}}^0$$

$$- [\ddot{\sigma}^0 (\bar{\delta}\bar{X})^2 + 2\bar{\delta}\bar{X}\bar{\delta}\dot{\sigma}^0 + \bar{\delta}^2 \sigma^0 + \dot{\sigma}^0 \bar{\sigma}^0]$$

$$+ \dot{\sigma}^0 \dot{\bar{\sigma}}^0 (X - \bar{X})\}, \tag{3.21}$$

where **A** and **B** take on values of 0 and 1. The $\tilde{\sigma}^0$ quantities are evaluated at the intersection of γ with μ and the σ^0 quantities at the intersection of $\bar{\gamma}$ with μ, i.e., $\tilde{\sigma}^0(u, \zeta, \bar{\zeta})$ and $\sigma^0(\bar{u}, \bar{\zeta}, \zeta)$. The dots denote $\partial/\partial u$ and the δ ($\bar{\delta}$) operator is applied to $\tilde{\sigma}^0$, $\dot{\tilde{\sigma}}^0$, and $\delta\tilde{\sigma}^0$ (σ^0, $\dot{\sigma}^0$, and $\bar{\delta}\sigma^0$) keeping u and $\bar{\zeta}$ ($\bar{u}$ and ζ) fixed. All other nonzero components of the curvature tensors can be obtained by using (1.10) and (1.11).[15] (3.21) shows that the components of the curvature tensor are combinations of quantities K, K_1, K_2, which have the striking form

$$K = \bar{\delta}^2 \sigma^0 + \dot{\sigma}^0 \tilde{\sigma}^0 - \delta^2 \tilde{\sigma}^0 - \dot{\tilde{\sigma}}^0 \sigma^0 - \delta^0 \dot{\tilde{\sigma}}^0 (X - \bar{X}),$$

$$K_1 = \delta \dot{\tilde{\sigma}}^0, \quad K_2 = -\ddot{\tilde{\sigma}}^0, \tag{3.22}$$

when $\Sigma = 0$, γ and $\bar{\gamma}$ intersect μ at the same point ($u = \bar{u}$), and K reduces to

$$K_0 = \bar{\delta}^2 \sigma^0 + \dot{\sigma}^0 \tilde{\sigma}^0 - \delta^2 \tilde{\sigma}^0 - \dot{\tilde{\sigma}}^0 \sigma^0. \tag{3.23}$$

The quantities K_2, K_1, K_0 are the respective R^{-1}, R^{-2}, and portion of the R^{-3} part of the gravitational (Weyl tensor) field of $\mathcal{M}$ in a standard notation (cf. Refs. 16, 17), i.e., ψ_4^0, ψ_3^0 and $\psi_2^0 - \bar{\psi}_2^0$, R being an affine parameter on outgoing null geodesics.

The components of the Kähler "Ricci tensor" can be obtained either from contraction of (3.21) or from the coordinate expression

$$K_{\mu'\nu} = K_{\mu'\lambda\nu}{}^{\lambda} = \frac{1}{2} \frac{\partial^2 \ln |G|}{\partial x^{\mu'} \partial x^{\nu}}, \tag{3.24}$$

where $|G|$ is the determinant of the matrix $G_{\Gamma\bar{\Lambda}}$. The Ricci tensor turns out to be identically zero.

As a final remark, we note that if the Kähler scalar Σ is replaced by $\log\Sigma$, then we obtain a Kähler metric defined on the *three-dimensional projective* asymptotic twistor space. This metric becomes singular, however, on $\Sigma = 0$. It has signature $(++-)$ on the region $\Sigma > 0$ and $(+--)$ on $\Sigma < 0$. In this form, the Kähler metric bears strong resemblance to the Bergmann metric[18] for bounded domains in C^n. This may have relevance particularly when $\mathcal{T}$ is the twistor space defining a nonlinear graviton,[19] since then the full extent of $\mathcal{T}$ is the region $\Sigma > 0$, with boundary $\Sigma = 0$. It seems likely that the metric studied here should also have significance for the study of such "gravitons."

*This research was supported in part by a grant from the National Science Foundation.

[1]R. Penrose and M. A. H. MacCallum, Phys. Repts. C 6, No. 4 (1973).
[2]R. Penrose, "Twistor theory, its aims and achievements," in *Quantum Gravity*, edited by C. J. Isham, R. Penrose, and D. W. Sciama (Oxford U. P., Oxford, 1975).
[3]S. Kobayashi and K. Nomizu, *Foundation of Differential Geometry* (Interscience, New York, 1963).
[4]J. A. Schouten, *Ricci-Calculus* (Springer-Verlag, Berlin, 1954).
[5]K. Yano, *Differential Geometry on Complex and Almost Complex Spaces* (MacMillan, New York, 1965).
[6]Thus K is $2n$-dimensional as a real manifold with a preferred set of complex coordinate charts related to one an-

other by holomorphic (i.e., complex analytic) coordinate transformations.

[7]Boldface types are used to denote ordinary numerical tensor component indices in order to distinguish them from the corresponding abstract indices.[8] Thus, $\alpha, \beta, \cdots, a, b, \cdots, \Gamma, \Theta, \cdots$ are the abstract versions of corresponding component indices $\boldsymbol{\alpha}, \boldsymbol{\beta}, \cdots, \mathbf{a}, \mathbf{b}, \cdots, \boldsymbol{\Gamma}, \boldsymbol{\Theta}, \cdots$.

[8]R. Penrose, "Structure of Space-Time," in *Battelle Rencontres*, edited by C.M. DeWitt and J.A. Wheeler (Benjamin, New York, 1968).

[9]There is complete symmetry (via the operation of complex conjugation) between the operators $\Box_\alpha$ and $\Box^\alpha$ even though the coordinate expressions (1.16) look very different from one another. The asymmetry in the coordinate expressions arises merely from the fact that the unprimed indices have been notationally singled out.

[10]For many purposes it is convenient to adopt a standard procedure of replacing "bars" over quantities by "tildes," so while a quantity $\bar{\xi}$ would initially have been the complex conjugate of ξ, upon complexification we would introduce a new quantity $\tilde{\xi}$ which is independent of ξ (cf. Sec. 3). With this notation w would be written $\tilde{z}$.

[11]The construction to be described will, of course, equally well refer to $\mathscr{I}^-$ as to $\mathscr{I}^+$. But for retarded gravitational fields the structure of $\mathscr{I}^-$ is identical to that for Minkowski space.

[12]R. Penrose, "Relativistic symmetry groups," in *Group Theory in Non-Linear Problems*, edited by A.O. Barut (Reidel, Dordrecht, 1972).

[13]R.O. Hansen and E.T. Newman (to be published).

[14]R. Penrose, Proc. Roy. Soc. London A **284**, 159 (1965).

[15]The curvature tensor of $\mathcal{CT}$ can be calculated in the same way starting from (3.12). The results are similar with all barred quantities replaced by tilded quantities.

[16]E.T. Newman and R. Penrose, J. Math. Phys. **3**, 566 (1962).

[17]E.T. Newman and T. Unti, J. Math. Phys. **3**, 891 (1962).

[18]S. Bergman, Math. Surveys, No. 5 (Amer. Math. Soc., Providence, R.I., 1968), 2nd ed.

[19]R. Penrose, General Relativity Gravitation **7**, 31 (1976).

85
THE TWISTOR PROGRAMME

Repts. Math. Phys. **12** (1977), no. 1, 65–76

A handy summary of basic twistor ideas.

THE TWISTOR PROGRAMME*

ROGER PENROSE

Mathematical Institute, Oxford, England

(Received October 5, 1976)

The formalism of twistors provides a new approach to the description of basic physics. The points of Minkowski space-time are represented by 2-dimensional linear subspaces of a complex 4-dimensional vector space (flat twistor space) on which a Hermitian form of signature $++--$ is defined. Free massless fields can be represented in terms of the sheaf cohomology of portions of this space. Twistor space (or a suitable part of it) can be expressed in two different ways as a complex fibration. If one or the other fibration structure is deformed, the resulting space represents not empty Minkowski space but, in one case, the general "right-flat" solution of Einstein's vacuum equations and, in the other, the general (left-handed) solution of Maxwell's equations. These provide the most primitive types of interaction (gravitational or electromagnetic) which may generalize to other fields in a comprehensive twistor scheme for the description of elementary particles.

Complex numbers are the basic ingredients of the mathematics of quantum physics. The linear superposition of quantum states is carried out with *complex* coefficients, these, (when normalized) being the so-called *probability amplitudes* of the theory. The normal mathematical description of space-time, on the other hand, uses only *real* numbers. Thus, the space-time continuum is taken to be a real differentiable manifold; and most physical quantities (e.g. velocity, energy, momentum, temperature), arising either directly or indirectly from space-time concepts, are normally also described in terms of real numbers. But physics is a unity; and it is to be hoped that eventually the whole of physics will be described in a unified way. If quantum physics and space-time geometry are to be unified, then the type of mathematical description appropriate to one ought also to apply to the other.

One of the main motivations behind twistor theory is, indeed, that it supplies a mathematical description of physics based in an essential way on a *complex* structure, where four-dimensional space-time geometry and quantum mechanical linear superposition both arise as intimately related aspects of this same complex twistor structure. Electromagnetic as well as gravitational interactions find natural descriptions in terms of two

* Presented on the Symposium on Methods of Differential Geometry in Physics and Mechanics, June 1976, Warsaw.

kinds of "deformation" of this basic twistor space structure. Further, a general scheme for the classification of leptons and hadrons can be given in terms of pairs or triples of twistors, respectively. It forms part of a programme for handling quantum mechanical interactions between particles, with a view to a comprehensive twistor version of quantum field theory.

An interrelation between the complex numbers of quantum mechanics and the structure of space arises already in the familiar quantum description of spin. The various pure states of (non-relativistic) spin $\frac{1}{2}$ correspond to the different possible directions in space (the possible directions of the spin-axis). Representing these directions by the points of a sphere S^2, we have a 1-1 relation between this S^2 and the rays in the 2-dimensional complex vector space of states. The S^2 may be viewed as the Riemann sphere (related by stereographic projection to the Argand plane) of ratios of the two complex components representing the state. We see that there is an intimate relation between the 3-dimensionality of space and the fact that it is *complex* numbers that are used in the quantum-mechanical superposition law. Put another way, this relation between the complex quantum structure and spatial structure finds mathematical expression in the 2-1 local isomorphism:

$$\mathrm{SU}(2) \to \mathrm{SO}(3).$$

Passing to (special) relativity, we re-interpret this S^2 as the space of (past or future) null directions in space-time, rather than directions in Euclidean 3-space. The above local isomorphism extends to its relativistic version:

$$\mathrm{SL}(2, \boldsymbol{C}) \to \mathrm{O}_+^\uparrow(1, 3)$$

(the right-hand side denoting the restricted Lorentz group). The group $\mathrm{SL}(2, \boldsymbol{C})$ induces transformations of the Riemann sphere to itself which are precisely those (the Möbius transformations) which preserve the structure of S^2 as a compact manifold of 1 complex dimension. This complex manifold has, by the above correspondence, a direct physical interpretation as the "celestial sphere" of an observer ([5]). The real space-time geometry of a null cone can thus be partially described in terms of a *complex* structure.

Twistor theory carries this correspondence further to enable the entire space-time geometry to be interpreted in complex manifold terms. The celestial sphere is concerned only with *directions* of photons. Let us consider, instead, a more complete space, each of whose points is to describe a right-handed "classical" spinning photon. The helicity is fixed at $s = +1$ (with $\hbar = 1$) and we obtain, as we shall see shortly, a space $\boldsymbol{PT}^+$ which is a manifold of 3 complex dimensions ([6], [9], [14]). (In real terms $\boldsymbol{PT}^+$ is 6-dimensional, two of these real dimensions corresponding to the direction of motion, one to the energy, and, roughly speaking, three to position.) This complex structure arises only because the spin is non-zero, the photon being characterized solely in terms of its classical momentum and angular momentum.

The space $\boldsymbol{PT}^+$ turns out to be describable as an open region in a complex projective 3-space $\boldsymbol{PT}$ $(= \boldsymbol{CP}_3)$. The boundary of $\boldsymbol{PT}^+$ in $\boldsymbol{PT}$ is a 5-real-dimensional manifold $\boldsymbol{PN}$

whose points may be taken to represent the null straight lines in (compactified) Minkowski space $\boldsymbol{M}$. (The photons are not themselves describable in terms of null straight lines in space-time. Only their high-energy limits are.) The complement of $\boldsymbol{PT}^+ \cup \boldsymbol{PN}$ in $\boldsymbol{PT}$ is another open region, $\boldsymbol{PT}^-$, whose points represent the left-handed (helicity $s = -1$) "classical" photons. The most direct correspondence between space-time and the complex manifolds $\boldsymbol{PT}^+$ and $\boldsymbol{PT}^-$ is achieved via their common boundary manifold $\boldsymbol{PN}$. In $\boldsymbol{M}$, any point may be represented by the system of null lines through it. This system corresponds to a locus in $\boldsymbol{PN}$ which is, a fact, a complex projective straight line. It turns out that *every* complex projective line in $\boldsymbol{PT}$ represents a point in the *complexified* space-time $\boldsymbol{CM}$, while those in $\boldsymbol{PN}$ represent *real* points in $\boldsymbol{M}$ (possibly at infinity). The lines lying entirely in $\boldsymbol{PT}^+$ represent points of $\boldsymbol{CM}$ whose imaginary parts are timelike and past-pointing (the forward tube); and correspondingly for $\boldsymbol{PT}^-$. The group of holomorphic self-transformations of $\boldsymbol{PT}^+$ (or, equivalently, of $\boldsymbol{PT}^-$) is the 15-parameter restricted conformal group $C^{\uparrow}_{+}(1, 3)$ on $\boldsymbol{M}$, this being 1-4 locally isomorphic with SU(2, 2) via the succession of 2-1 local isomorphisms

$$\mathrm{SU}(2, 2) \to \mathrm{O}^{\uparrow}_{+}(2, 4) \to \mathrm{C}^{\uparrow}_{+}(1, 3).$$

The study of complex manifolds is not greatly familiar to theoretical physicists. They work primarily with real manifolds, and the necessary mathematical machinery ([1], [3], [19]) is not easy to assimilate. It is also necessary to build up new intuitions appropriate to the subject. Often there is a *rigidity* present with complex manifolds and holomorphic functions, which would be totally absent in the case of real manifolds and functions. But there can be flexibility too. Very roughly: one expects some rigidity in those directions in where there is compactness and perhaps flexibility in directions where there is non-compactness. The contrast with the real case is profound. For example, any real closed curve on a real manifold can be continuously displaced in many ways, the displaced curves constituting an infinite-dimensional system. (This fact is independent of the differentiability class of the curve and manifold. It holds just as well for C^{ω} as for C^{∞}, C^1 or C^0.) However, a closed holomorphic complex curve (which in real terms would be a closed surface) on a fixed complex manifold can be displaced only within finite-dimensional holomorphic system. In fact, in some cases the curve cannot even be displaced at all (0-dimensional system).

Again, consider smooth transformations of a compact manifold to itself. In the real case we always get an infinite-parameter system and in the complex case a finite- (possibly 0-) parameter system. Indeed, we have just noted one instance of the latter, whereby the 6-parameter Lorentz group was realized as the holomorphic self-transformations of the Riemann sphere. Related to this is the question of rigidity of complex structure. A continuous (or holomorphic) connected family of complex manifolds may be such that any two members of the family are necessarily complex-diffeomorphic (as would be the case if one of the members were $\boldsymbol{PT}$); or, in the general case where the manifolds are compact, a discrete or finite-parameter system of complex-diffeomorphically distinct manifolds:

or, in the non-compact case, sometimes an infinite-parameter system (as could be the case if one member were PT^+). The fact that an infinite number of parameters of information can be "stored" in the *global* structure of a (non-compact) *complex* manifold—whereas locally, the only information would be the dimension of the manifold—is of crucial significance in twistor interaction theory.

Consideration of the above questions leads us into *sheaf cohomology theory*, which provides us with the beginnings of a complex differential topology. (Roughly speaking, sheaf theory provides the "linearized topology" of small perturbations.) It turns out that (holomorphic) sheaf theory is also important in the twistor formalism for other reasons. A brief description of the relevant theory ([1], [3], [19]) follows.

Let X be a (Hausdorff, paracompact) complex manifold and let $\mathscr{U} = \{U_i\}$ be a locally finite covering of X by open sets. To specify a *sheaf* $\mathscr{S}$ we single out a (suitable) class of functions defined on open subsets of X. To fix ideas, we may take these functions simply to be arbitrary holomorphic functions on the subset. But more generally, the functions, could be holomorphic cross-sections of complex vector bundles. (And more specially, they could be restricted to be constants. Then the following construction specializes to standard Čech cohomology theory of "ordinary" topology.) To define a 0-*cochain* (with respect to $\mathscr{U}$ and with coefficients in $\mathscr{S}$) we select one function f_i (belonging to the chosen class) defined on each U_i of the covering; to define a 1-*cochain* we select one function f_{ij} defined on $U_i \cap U_j$, for each ordered pair U_i, U_j in the covering, subject to the restriction $f_{ij} = -f_{ji}$; to define a 2-*cochain* we select one function f_{ijk} defined on $U_i \cap U_j \cap U_k$, for each ordered triplet U_i, U_j, U_k, with the restriction that the f_{ijk} are skew in their indices; etc. The *coboundary operator* δ takes a p-cochain $\alpha = \{f_{i\ldots k}\}$ to a $(p+1)$-cochain $\delta\alpha = \{g_{i\ldots kl}\}$, where if $p = 0$ we have $g_{ij} = f_j - f_i$, if $p = 1$ we have $g_{ijk} = f_{jk} - f_{ik} + f_{ij}$, if $p = 2$ we have $g_{ijkl} = f_{jkl} - f_{ikl} + f_{ijl} - f_{ijk}$, etc. A p-chain α is a cocycle if $\delta\alpha = 0$ and a coboundary if α has the form $\alpha = \delta\beta$. (All coboundaries are cocycles since $\delta^2 = 0$.) Defining $H^p_{\mathscr{U}}(X, \mathscr{S})$ to be the complex vector space of p-cocycles *modulo* the p-coboundaries, we take the direct limit for finer and finer coverings $\mathscr{U}$ of X to obtain $H^p(X, \mathscr{S})$, the *p-th cohomology group of X with coefficients in $\mathscr{S}$*. (In practice it is not necessary to take this limit, as there are ways of ensuring that a *particular* choice of $\mathscr{U}$ is fine enough.)

To show how these ideas relate to twistor theory, I shall first briefly review the basic twistor formalism ([6], [9], [14]). A twistor Z^α, with components $(Z^0, Z^1, Z^2, Z^3) \in \boldsymbol{C}^4$, is an element of a 4-dimensional complex vector space $\boldsymbol{T}$. Its initial physical interpretation is that it describes the momentum p_a and angular momentum M^{ab} of a massless particle in flat space-time according to $Z^0 = \omega^0$, $Z^1 = \omega^1$, $Z^2 = \pi_{0'}$, $Z^3 = \pi_{1'}$, where

$$p_{AA'} = \bar{\pi}_A \pi_{A'}, \qquad M^{AA'BB'} = i\omega^{(A}\bar{\pi}^{B)}\varepsilon^{A'B'} - i\varepsilon^{AB}\bar{\omega}^{(A'}\pi^{B')}$$

(round brackets denoting symmetrization, where a standard 2-spinor notation is being employed for the capital Latin indices, relating $V^{AA'}$ to V^a and to $V_{AA'}$ by $V^{00'} = V_{11'} = 2^{-1/2}(V^0 + V^1)$, $V^{01'} = -V_{10'} = 2^{-1/2}(V^2 + iV^3)$, $V^{10'} = -V_{01'} = 2^{-1/2}(V^2 - iV^3)$, $V^{11'} = V_{00'} = 2^{-1/2}(V^0 - V^1)$, the space-time signature being $+---$). The complex

conjugate twistor $\bar{Z}_\alpha$ belongs to the dual space T^* to T, the standard components of $\bar{Z}_\alpha$ being related to those of Z^α by $\bar{Z}_0 = \overline{Z^2}$, $\bar{Z}_1 = \overline{Z^3}$, $\bar{Z}_2 = \overline{Z^0}$, $\bar{Z}_3 = \overline{Z^1}$. We write

$$Z^\alpha = (\omega^A, \pi_{A'}), \quad \bar{Z}_\alpha = (\bar{\pi}_A, \bar{\omega}^{A'}),$$

and

$$Z^\alpha \bar{Z}_\alpha = \omega^A \bar{\pi}_A + \pi_{A'} \bar{\omega}^{A'} = 2s,$$

finding that s is the *helicity* of the massless particle, defined by $S_a = sp_a$, where $S_a = \frac{1}{2}\epsilon_{abcd} p^b M^{cd}$ is the Pauli–Lubanski spin vector. We note that ω^A depends on the choice of origin-point, whereas $\pi_{A'}$ does not.

To obtain the correspondence between twistors and space-time geometry, we define the relation:

$$\textit{the point } x \textit{ lies on the twistor } Z^\alpha \Leftrightarrow \omega^A = ix^{AA'}\pi_{A'},$$

where x^a is the position vector of $x \in \boldsymbol{CM}$ with respect to the chosen origin. Only when $s = 0$ (i.e. Z^α is a null twistor: $Z^\alpha \bar{Z}_\alpha = 0$) can we find real points on Z^α, in which case the locus of such x is a null straight line in $\boldsymbol{M}$ (assuming $Z^\alpha \neq 0$). For general Z^α, the locus of complex points on Z^α is a totally null (i.e. with only null tangent vectors), totally geodesic complex 2-surface in $\boldsymbol{CM}$, the tangent vectors having the form $\lambda^A \pi^{A'}$ for fixed $\pi^{A'}$ and varying λ^A. A dual twistor W_α correspondingly defines a totally null 2-surface of the complex conjugate type (fixed λ^A and varying $\pi^{A'}$). The twistor orthogonality relation $Z^\alpha W_\alpha = 0$ states that the two totally null surfaces have a (complex) point in common.

The space $\boldsymbol{T}$ is divided into three disjoint regions $\boldsymbol{T}^-$, $\boldsymbol{N}$, $\boldsymbol{T}^+$, defined according as $Z^\alpha \bar{Z}_\alpha$ is negative, zero, or positive. *Projective* twistor space $\boldsymbol{PT}$ is the set of equivalence classes $\{\lambda Z^\alpha | \ \lambda \in \boldsymbol{C}, \lambda \neq 0\}$ with $Z^\alpha \neq 0$. It has the structure of $\boldsymbol{CP}_3$ and can be divided into three corresponding regions $\boldsymbol{PT}^-$, $\boldsymbol{PN}$, $\boldsymbol{PT}^+$. Note that "phase" transformations $Z^\alpha \to e^{i\theta} Z^\alpha$ (θ real) leave p_a and M^{ab} invariant whereas real scalings of Z^α do not. To say that Z^α describes a right-handed "classical photon" would be to fix the real scaling by $2s = Z^\alpha \bar{Z}_\alpha = 2$ (choosing units so that $\hbar = 1$). Thus, the space $\boldsymbol{PT}^+$ indeed represents such right-handed photons, as mentioned earlier, and $\boldsymbol{PT}^-$ similarly represents left-handed photons ($Z^\alpha Z_\alpha = -2$). The various above-mentioned properties of these spaces are now readily derived, the group SU(2, 2) arising because the Hermitian form $Z^\alpha \bar{Z}_\alpha$ has signature $++--$.

Next, a (first) quantization scheme is introduced whereby the Z^α and $\bar{Z}_\alpha$ become *operators* subject to commutation rules

$$[Z^\alpha, Z^\beta] = 0, \quad [Z^\alpha, \bar{Z}_\beta] = \delta^\alpha_\beta, \quad [\bar{Z}_\alpha, \bar{Z}_\beta] = 0.$$

The motivation for this choice comes from various different directions. The choice is compatible with the normal position-momentum canonical quantization, but seems to involve an additional "twistorial" feature expressing conjugacy between helicity and the twistor phase. There is a compatibility with the conformal invariance of massless fields and also with certain general-relativistic considerations ([7]). These features are

related to the fact that, classically, $i d Z^\alpha \wedge d\bar{Z}_\alpha$ is the appropriate invariant symplectic form on $\boldsymbol{T}$. The choice of twistor variables Z^α means that we have, in effect, selected a *complex polarization* as the basis of the quantization procedure ([20]). This differs from the alternative approach of Souriau [18], which uses a real polarization for massless particles and leads, in effect, to selecting $(\bar{\pi}_A, \pi_{A'})$ as variables rather than $Z^\alpha = (\omega^A, \pi_{A'})$. The twistor approach has the advantage of manifest conformal invariance, but the fact that we now consider *holomorphic* functions $f(Z^\alpha)$ for our wave functions (since "independence of $\bar{Z}_\alpha$" for f is the holomorphic condition $\delta f/\partial \bar{Z}_\alpha = 0$) leads us inevitably into sheaf cohomology theory.

The operator $\bar{Z}_\alpha$, acting on $f(Z^\alpha)$, is now represented by $-\partial/\partial Z^\alpha$. Substituting this into the expressions for p_a and M^{ab} (i.e. replacing $\bar{\omega}^{A'}$ by $-\partial/\partial \pi_{A'}$ and $\bar{\pi}_A$ by $-\partial/\partial \omega^A$) we obtain differential operators which satisfy the same commutation rules as the normal ones, where p_a and M_{ab} stand for the infinitesimal generators of translations and Lorentz rotations, respectively, for the Poincaré group. No factor ordering problems arise for the twistor-operator definitions of these quantities. The helicity operator $\hat{s}$ needs more care, but the result is unambiguous:

$$\hat{s} = \tfrac{1}{4}(Z^\alpha \bar{Z}_\alpha + \bar{Z}_\alpha Z^\alpha) = -\tfrac{1}{2} Z^\alpha \frac{\partial}{\partial Z^\alpha} - 1 .$$

This implies that the twistor wave function $f(Z^\alpha)$ for a massless particle of helicity s must be *homogeneous of degree* $-2s-2$.

The holomorphic properties of f imply that its domain cannot be the whole of $\boldsymbol{T}$, nor even the whole of $\boldsymbol{T}^+$. However, we can interpret f in a different way, namely as providing a representative 1-cocycle defining an element of a 1st cohomology group of complex sheaf cohomology. The homogeneous nature of f entails that it can be regarded as a "twisted function" (i.e. element of a complex line bundle) on a portion of $\boldsymbol{PT}$. The (holomorphic) twisted functions corresponding to homogeneity n give us a sheaf over $\boldsymbol{PT}$ which will be denoted by $\mathcal{O}(n)$. Now in simple cases we can use *two* open sets U_1 and U_2 to cover $\boldsymbol{PT}^+$, the overlap $U_1 \cap U_2$ being a kind of annular region on which $f = f_{12} = -f_{21}$ can be defined, the singularities of f falling into two disjoint regions on $\boldsymbol{PT}^+$, namely $U_1 - U_2$ and $U_2 - U_1$. In more general cases we might need more open sets, and correspondingly more functions. But in any case, we find ([11]) that a valid twistor description for the wave function of a helicity s massless particle is an element of the sheaf cohomology group

$$H^1(\overline{\boldsymbol{PT}^+}, \mathcal{O}(-2s-2)).$$

The use of $\boldsymbol{PT}^+$ automatically ensures a positive frequency condition; the extension to its closure $\overline{\boldsymbol{PT}^+} = \boldsymbol{PT}^+ \cup \boldsymbol{PN}$ ensures that the wave function is normalizable. A covering of $\overline{\boldsymbol{PT}^+}$ by just two closed sets is actually sufficient here, but we can also generalize to situations in which $\overline{\boldsymbol{PT}^+}$ is replaced by some other subset X of $\boldsymbol{PT}$. Then a more complicated covering may be needed.

A transcription to space-time terms is achieved explicitly by means of one of the following two contour integrals ([8], [9], [10], [14]) (assuming that we are in the simple situation in which only one f is required—otherwise a certain generalization involving "branched" contour integrals ([11]) would be needed):

$$\underbrace{\phi_{AB\ldots L}}_{-2s} = \frac{1}{(2\pi i)^2}\oint \frac{\partial}{\partial\omega^A}\frac{\partial}{\partial\omega^B}\cdots\frac{\partial}{\partial\omega^L} f(Z^\alpha)\,d^2\pi \qquad (s = 0, -\tfrac{1}{2}, -1, \ldots),$$

$$\underbrace{\phi_{A'B'\ldots L'}}_{2s} = \frac{1}{(2\pi i)^2}\oint \pi_{A'}\pi_{B'}\cdots\pi_{L'} f(Z^\alpha)\,d^2\pi \qquad (s = 0, \tfrac{1}{2}, 1, \ldots).$$

where

$$Z^\alpha = (\omega^A, \pi_{A'}) = (ix^{AA'}\pi_{A'}, \pi_{A'}), \qquad d^2\pi = d\pi_{0'}\wedge d\pi_{1'}.$$

The integral is over a closed 2-dimensional contour in $\pi_{A'}$-space for each fixed $x^{AA'}$ and yields the value of the field $\phi_{\ldots}$ at the point x. The contour moves continuously as x is varied. By virtue of the homogeneity of f, the integral can also be simply re-expressed as over a 1-dimensional contour, where $2\pi i(\pi_{0'}d\pi_{1'} - \pi_{1'}d\pi_{0'})$ replaces $d^2\pi$. As a consequence of these expressions, the massless free-field equations are automatically satisfied:

$$\Box\phi = 0 \ (s = 0), \qquad \nabla^{AA'}\phi_{AB\ldots L} = 0 \ (s < 0), \qquad \nabla^{AA'}\phi_{A'B'\ldots L'} = 0 \ (s > 0).$$

For $s = \pm\frac{1}{2}$ this gives the Weyl neutrino equation; for $s = \pm 1$, the 2-spinor version of Maxwell's free-field equations; for $s = \pm 2$, the gauge-invariant 2-spinor form of the linearized Einstein free field.

Note that to add two space-time fields (wave functions) we simply add the corresponding cohomology group elements, but this need *not* be expressible as just adding the corresponding f's. This is because the singularity structure of the two functions need not be compatible. Another apparent anomaly is the asymmetry between right-handed ($s > 0$) and left-handed ($s < 0$) helicities. The answer to this, at one level of understanding, is that symmetry is restored if we allow ourselves to pass to the dual twistor space $\boldsymbol{T}^*$ and re-express the integral in terms of $\bar{Z}_\alpha$. This involves applying a notion of "twistor transform" to the functions f, but I shall not discuss the details here. But at a deeper level lies the question of whether such symmetry is desirable in any case. Arguments can be presented ([13]) that CPT-invariance should not be completely true for all interactions of Nature. If this proves to be the case, then twistor theory provides a possible formalism for handling the situation. Either twistor space $\boldsymbol{T}$, or its dual $\boldsymbol{T}^*$, would have to be taken as more primitive than the other.

We come now to the *active* role played by twistor functions in generating interactions. As mentioned earlier, sheaf cohomology theory can be used to express the infinitesimal (or "linearized") deformations of complex spaces. The exponentiation of such deformations can yield a genuine change in structure. Consider first the case where $f(Z^\alpha)$ is homogeneous of degree zero (negative helicity photon: $s = -1$). We regard $\boldsymbol{T}^+$ as a complex line bundle over $\boldsymbol{PT}^+$ and use f to deform this bundle ([2]) to obtain a curved twistor

structure $\mathscr{T}^+$. For simplicity assume that $\boldsymbol{T}^+$ is covered by just two regions U_1 and U_2 with f defined on $U_1 \cap U_2$, where each of U_1, U_2 is a union of fibres. (Thus, $\boldsymbol{PT}^+$ is a union of two regions $\boldsymbol{PU}_1$ and $\boldsymbol{PU}_2$ whose respective lifts back into the bundle are U_1 and U_2.) We now piece U_1 and U_2 together in a new way to obtain the space $\mathscr{T}^+$: take $\hat{Z}^\alpha \in U_1$ and $Z^\alpha \in U_2$, and define the transition functions on the overlap region on $\mathscr{T}^+$ by

$$\hat{Z}^\alpha = e^{f(Z^\alpha)} Z^\alpha.$$

This gives $\mathscr{T}^+$ the structure of a fibre bundle over $\boldsymbol{PT}^+$. (The fact that f is homogeneous of degree zero is necessary for this, since $Z^\alpha \to \lambda Z^\alpha$ must be accompanied by $\hat{Z}^\alpha \to \lambda \hat{Z}^\alpha$.)

Since the relation between twistor space structure and space-time geometry can be achieved merely by referring to the *projective* twistor space structure, the space-time geometry arising from $\mathscr{T}^+$ remains the same as that of $\boldsymbol{T}^+$, i.e. Minkowskian. However, the "phases" of the twistors are affected in a way which when interpreted in space-time terms, leads to the type of "connection" for $\boldsymbol{M}$ induced by an electromagnetic potential. This turns out to correspond to the same electromagnetic field (or wave function) as the one obtained by the contour integral, but now can be thought of as describing the photon in an *active* capacity, since the description incorporates into the structure of the twistor space the way that a photon interacts. (The construction is due to R. Ward [2], incorporating some ideas of G. A. J. Sparling.)

The use of the portion $\boldsymbol{PT}^+$, of $\boldsymbol{PT}$, as base space in the above construction arises because we consider a positive-frequency free photon. However, the construction can also be used when the base space is some other suitable portion X of $\boldsymbol{PT}$, provided that we suitably generalize the above to the case when X is covered by more than two patches. If X contains "holes" corresponding to world-tubes in $\boldsymbol{M}$ representing charged matter, then the above construction can be carried out only if the charge of each hole is an integral multiple of some fixed value ([2], [11]).

A similar construction can also be given in the gravitational case ([12]), where f has homogeneity degree 2 (i.e. helicity -2). Now the "unperturbed" $\boldsymbol{T}^+$ is to be viewed as a "bundle" in a different way, the base space being the $\pi_{A'}$-spin-space and the projection map being $(\omega^A, \pi_{A'}) \mapsto \pi_{A'}$. (Although this topologically assigns a bundle structure to $\boldsymbol{T}^+$, it does not, strictly speaking, assign one globally holomorphically. We refer to this structure as a holomorphic fibration.) We wish to "deform" $\boldsymbol{T}^+$ suitably, but it will be easier at this stage if we do not concern ourselves too much with the global structure of the resulting space $\mathscr{T}$. For simplicity assume that $\mathscr{T}$ can be constructed from two open regions U_1, $U_2 \subset \boldsymbol{T}$, with f defined on the overlap $U_1 \cap U_2$. We piece U_1 and U_2 together in a different way to obtain $\mathscr{T}$, where if $\hat{Z}^\alpha \in U_1$ and $Z^\alpha \in U_2$, now we have the transition relation

$$\hat{Z}^\alpha = \exp\left\{\varepsilon^{AB} \frac{\partial f(Z^\alpha)}{\partial \omega^A} \frac{\partial}{\partial \bar{\omega}^B}\right\} Z^\alpha$$

on the overlap. (Note the appearance of the exponentiation of an *operator*; a generating

function method is also possible ([12]), in which this is avoided.) The form of this relation is such as to guarantee that the complex symplectic structure $\varepsilon^{AB}\partial^2/\partial\omega^A\partial\omega^B$ is preserved on each fibre and so is the Euler operator $Z^\alpha\partial/\partial Z^\alpha$. The base space is to be the whole $\pi_{A'}$-spin-space, and we obtain a holomorphic fibration as before.

To find the space-time interpretation of $\mathscr{T}$, we need to define "space-time points". These are given by the *holomorphic cross-sections* of $\mathscr{T}$, the space of these cross-sections forming a 4-complex dimensional manifold denoted by $\mathscr{M}$. Two points of $\mathscr{M}$ are defined to be *null-separated* if the corresponding sections of $\mathscr{T}$ *intersect*. It is possible to extend this definition of complex conformal structure on $\mathscr{M}$ to a full complex Riemannian metric g_{ab} on $\mathscr{M}$. It turns out that this metric *automatically satisfies Einstein's vacuum equations* $R_{ab} = 0$. But, in addition, the self-dual ("right-handed") part of the Weyl curvature also vanishes:

$$C^+_{abcd} := \tfrac{1}{2}(C_{abcd} - iC^*_{abcd}) = 0;$$

whereas the anti-self-dual ("left-handed") part generally does not:

$$C^-_{abcd} := \tfrac{1}{2}(C_{abcd} + iC^*_{abcd}) \neq 0.$$

Such "space-times" (necessarily complex—or else with even signature) are called *right flat* ([12]).

We note the spinor representations of these tensors:

$$C^+_{AA'BB'CC'DD'} = \varepsilon_{AB}\varepsilon_{CD}\tilde{\Psi}_{A'B'C'D'},$$
$$C^-_{AA'BB'CC'DD'} = \Psi_{ABCD}\varepsilon_{A'B'}\varepsilon_{C'D'},$$

where each of $\tilde{\Psi}_{A'B'C'D'}$ and Ψ_{ABCD} is symmetric; and they are independent in the complex case. For a real space-time we have

$$\tilde{\Psi}_{A'B'C'D'} = \bar{\Psi}_{A'B'C'D'}$$

whereas for a right- (left-) flat space-time

$$\tilde{\Psi}_{A'B'C'D'} = 0 \quad (\Psi_{ABCD} = 0).$$

If $\mathscr{T}$ has the appropriate global properties to be suitably analogous to T^+, then it describes a left-handed *non-linear graviton*. In the weak-field limit, this agrees with the conventional definition for the wave function of a negative helicity graviton in terms of a spin 2 massless representation of the Poincaré group. But here the graviton is self-interacting in a way compatible with the non-linearities and curvature of Einstein's theory ([12]). (This particular approach to the non-linear graviton, using the above exponentiated operator expression, has been instigated by Sparling [2], and he has been able to show that the curvature in the weak field limit agrees with that given by the contour integral expression using the same f, i.e. $\Psi_{ABCD} = \phi_{ABCD}$. The general twistor-space construction of right-flat space-times evolved out of earlier work by E. T. Newman [4] concerning asymptotically flat space-times and their related "$\mathscr{H}$-spaces".)

Note that in the weak-field limit it is possible to take the real part of a complex field and still satisfy the linearized massless spin 2 equation. The above construction, in fact, yields the general solution. In the non-linear case we still have as much freedom as this

in our construction of right-flat complex space-times, but no means has emerged, as yet, for "taking the real part". If such could be found, then we should have a means of constructing the general solution of Einstein's vacuum equations!

In order to proceed further with twistor theory it appears to be necessary, for this and other reasons, to consider products of twistor spaces, e.g. $T\times T^*$ and $T\times T\times T$, and to examine the sheaf cohomology and deformation theory of suitable portions of these product spaces. In the case of Einstein's equations it seems that we are led most naturally to consider deformations of portions of $T\times T^*$, or of the space $CN = \{Z^\alpha, W_\alpha |\ Z^\alpha W_\alpha = 0\} \subset T\times T^*$, since it is possible to code the information of at least the conformal structure of a *general* space-time into the global complex structure of such a space ([9]). The product spaces $T\times T$ and $T\times T\times T$ arise in a general twistor scheme ([10]) for describing leptons and hadrons, respectively (although it is not yet completely clear whether the dual space T^* should be used in place of one or more of these T's—it makes no immediate difference). This scheme will next be described.

The central idea is that whereas functions of a single twistor $f(Z^\alpha)$ can be used to describe massless particles and fields, holomorphic functions of several twistors $F(\underset{1}{Z^\alpha}, \ldots, \underset{n}{Z^\alpha})$ can, in a similar way, be used to describe massive particles of various kinds. More correctly, we should not really be concerned specifically with such functions F but with a sheaf cohomology group element that F serves to define. Thus, in the first instance, we are concerned with the sheaf cohomology of suitable portions of $T\times T$ or of $T\times T\times T$, say, in order to obtain a twistorial description of free massive particles. At a later stage, we may expect that finite deformations of these spaces would play a role in expressing the interactions between the particles. However, at present not much is known about this, so I shall discuss only the twistor description of free particles.

First we require the many-twistor expressions ([9], [10], [14]) for momentum and angular momentum:

$$P_{AA'} = \sum_{r=1}^{n} \underset{r}{\bar\pi_A}\,\underset{r}{\pi_{A'}} = -\sum_r \underset{r}{\pi_{A'}}\,\partial/\partial\underset{r}{\omega^A},$$

$$M^{AA'BB'} = \sum_{r=1}^{n} \{i\underset{r}{\omega^{(A}}\,\underset{r}{\bar\pi^{B)}}\varepsilon^{A'B'} - i\varepsilon^{AB}\,\underset{r}{\bar\omega^{(A'}}\,\underset{r}{\pi^{B')}}\}$$

$$= \sum_r \{i\varepsilon^{A'B'}\,\underset{r}{\omega^{(A}}\,\partial/\partial\underset{r}{\omega_{B)}} + i\varepsilon^{AB}\,\underset{r}{\pi^{(A'}}\,\partial/\partial\underset{r}{\pi_{B')}}\},$$

where

$$\underset{r}{Z^\alpha} = (\underset{r}{\omega^A}, \underset{r}{\pi_{A'}}) = (i\underset{r}{x^{AA'}}\,\underset{r}{\pi_{A'}}, \underset{r}{\pi_{A'}}).$$

We now have $m^2 = P_{AA'}P^{AA'}$ ($\geqslant 0$, in unitary descriptions). For suitable F's, we can generate space-time fields (wave functions) by means of contour integrals:

$$\Phi_{A'\ldots C\ldots E'\ldots H}(x) = \frac{1}{(2\pi i)^{2n}} \oint \underset{1}{\pi_{A'}} \ldots \frac{\partial}{\partial\underset{1}{\omega^C}} \ldots \underset{r}{\pi_{E'}} \ldots \frac{\partial}{\partial\underset{n}{\omega^H}} F(\underset{r}{Z^\alpha})\, d^2\underset{1}{\pi} \wedge \ldots \wedge d^2\underset{n}{\pi}.$$

If F is unrestricted, then many such fields can be generated, depending upon which π-terms or $\partial/\partial\omega$-terms are inserted in the integrand. However, requiring F to be a simultaneous eigenstate of a certain complete set of commuting operators will ensure that *at most one* of these fields is non-zero and, furthermore, that this field is irreducible with respect to the Poincaré group. Thus, F itself (or the sheaf cohomology element it describes) can be used to generate a particle wave function, but it contains *more* information (quantum numbers) than such a wave function, namely the various eigenvalues of the operators. (Sometimes these eigenvalues are such that no space-time field description is possible at all—a significant observation due to Sparling—and our present view is that these F's describe higher particle resonances.)

The required complete set of commuting operators belongs to the enveloping algebra of the infinitesimal "internal" twistor group of (quantized) linear transformations on Z^α and $\overline{Z}_\alpha$ which preserves p_a and M^{ab}. Among these are the spin and mass operators, which are unambiguously identifiable, but also there are various other operators (with apparently only discrete spectra of values) defining extra internal quantum numbers.

The *twistor particle hypothesis* ([10]) is that these are *all* the quantum numbers possessed by the elementary particles of Nature, i.e. that F (or its cohomology class) affords a complete description of a particle, both as regards its external and internal parameters. I shall not enter into the detailed discussion of this here (which rests largely on work by Z. Perjés, [15], [16], [17], L. P. Hughston [2] and G. A. J. Sparling [17]), but mention only the broad features. First there are the 1-twistor particles. These are massless (described by $f(Z^\alpha)$, above) and have the "internal" group $U(1)$ (given by $Z^\alpha \to e^{i\theta}Z^\alpha$), providing the helicity quantum number. Second, we have 2-twistor particles, and we hypothesize that these are leptons ([2]). The "internal" group is a two-fold covering of $O(3)\times$ $\times E(2)$ (where $E(2)$ is the Euclidean group on the plane). There are two unidentified internal quantum numbers—related to charge, electron and muon lepton numbers. Third we have 3-twistor particles, which we suggest are hadrons ([2], [15], [16], [17]). The "internal" group may be taken ([15]) to be an "inhomogeneous SU(3)" with quantum numbers that can be related to those of standard SU(3) theory. In summary:

		external		scalar
T: massless particles	— 4 commuting operators	3	1	0
$T\times T$: leptons?	— 8 commuting operators	4	2	2
$T\times T\times T$: hadrons?	— 12 commuting operators	4	2	6
$T\times T\times T\times T$: J/Ψ???	— 16 commuting operators	4	2	10
		spatial momentum, spin in z-direction	m, J	internal

An additional feature of this scheme appears to be required; namely that any n-twistor particle ought also to be expressible as an $(n+1)$-twistor particle, possibly with some

"mixing" occurring between the two sets of twistors (related to the Cabibbo angle). This is the general problem of "twistor growing" and one promising approach towards this is provided by sheaf cohomology theory ([11]).

A related problem is how to describe many-particle states twistorially. Again, functions of several twistors are needed, and the expressions for p_a and M^{ab} must be the same as before. But the contour integral expressions for the wave functions are not quite the same as for one-particle states, suggesting that *different* sheaf cohomology groups of portions of $\boldsymbol{T}\times\boldsymbol{T}\times\ \ldots\ \times\boldsymbol{T}$ are involved for many-particle states from those for one-particle states. It now seems clear that a comprehensive twistor treatment of quantum field theory (which up until now has proceeded in a piecemeal fashion involving a few explicit contour integrals) must rest heavily on complex sheaf cohomology theory.

REFERENCES

[1] Gunning, R. C., H. Rossi: *Analytic functions of several complex variables*, Prentice Hall, Englewood Cliffs, N. J., 1965.

[2] Hughston, L. P., R. Penrose, G. A. Sparling, R. Ward: in *Twistor Newsletter, No.* **1**, preprint, Mathematical Institute, Oxford 1976.

[3] Morrow, J., K. Kodaira: *Complex manifolds*, Holt, Rinehart and Winston, New York 1971.

[4] Newman, E. T.: *Gen. Rel. Grav.* **7** (1976), 107.

[5] Penrose, R.: *Proc. Camb. Phil. Soc.* **55** (1959), 137.

[6] —: *J. Math. Phys.* **8** (1967), 345.

[7] —: *Int. J. Theor. Phys.* **1** (1968), 61.

[8] —: *J. Math. Phys.* **10** (1969), 38.

[9] —: in: *Quantum gravity*, eds. C. J. Isham, R. Penrose, D. W. Sciama, Oxford Univ. Press, 1975.

[10] —: in: *Quantum theory and the structures of time and space*, eds. L. Castell, M. Drieschner, C. F. von Weiszacker, C. Hanser, München 1975.

[11] —: in *Twistor Newsletter, No.* **2**, preprint, Mathematical Institute, Oxford 1976.

[12] —: *Gen. Rel. Grav.* **7** (1976), 31, 171.

[13] —: article in volume in honour of S. Chandrasekhar, ed. N. Lebovitz, Chicago Univ. Press, 1977.

[14] —, M. A. H. MacCallum: *Phys. Reports* **6C** (1973), 241.

[15] Perjés, Z.: *Rev.* **11D** (1975), 2031.

[16] —: *Reports Math. Phys.* this volume.

[17] —, G. A. J. Sparling: preprint 1976, to appear.

[18] Souriau, J-M.: *Structures des systèmes dynamiques*, Dunod, Paris 1970.

[19] Wells, R. O. Jr.: *Differential analysis on complex manifolds*, Prentice Hall, Englewood Cliffs, N. J., 1973.

[20] Woodhouse, N. M. J.: in *Group theoretical methods in physics*, Springer-Verlag, Berlin 1976.

86
SPACE-TIME SINGULARITIES

Proceedings of the first Marcel Grossman meeting on general relativity. Ed. R. Ruffini. ICTP Trieste North-Holland, Amsterdam, 1977

An early mention of the Weyl curvature hypothesis is on p. [178].

128 Space-time singularities; *Proceedings of the first Marcel Grossman meeting on general relativity.* Ed. R. Ruffini. ICTP Trieste North-Holland, Amsterdam, 1977

SPACE-TIME SINGULARITIES

Roger Penrose

University of Oxford, England

ABSTRACT

The "singularity theorems" of classical general relativity imply the existence of singularities in space-time (i.e. places where the normal classical space-time descriptions must break down) whenever too large concentrations of mass are present, provided that reasonable requirements of causality and positivity of energy are satisfied. Such singularities may be of the initial exploding big-bang type or the result of final gravitational collapse inside a black hole--or they might possibly be naked singularities resulting from a collapse. Most relativists apparently believe that naked singularities will not occur in realistic collapse situations according to classical general relativity, but a proof is not yet forthcoming.

A convenient way to represent singularities is by the construction of "ideal points" ("TIP's" and "TIF's"). Initial, final and naked singularities may be characterized using this construction and their structure may be examined in detail.

A crucial difference between the geometry of initial and of final singularities seems probable, whereby the conformal curvature vanishes initially, but dominates over Ricci curvature finally. This may plausibly be related to the statistical time's arrow, and, more speculatively, to the minute time-asymmetry of local physical laws.

The existence of singularities in the structure of space-time is strongly indicated from both theory and observation. By such a singularity is meant a region of the universe at which our normal physical space-time picture breaks down. So to describe the nature of such a region in detail a new physical theory would be needed. According to a commonly expressed view, in the neighbourhood of a singularity, the radii of space-time curvature might be approaching 10^{-33} cm - the Planck length, at which quantum fluctuations in the metric ought in any case to render the normal Riemannian manifold picture of space-time inappropriate. Accordingly, the required new theory would be some satisfactory fully quantized version of Einstein's general relativity. Alternatively, it might turn out that when the radii of space-time curvature approach only the much more modest value of, say, 10^{-13} cm, a new theory would already be needed. Indeed, one could not expect much of present-day elementary particle theory to survive at a regime in which the approximation of a flat space-time ceases to be appropriate and the Poincaré group accordingly has no local relevance, various characteristic lengths involved in elementary particles being already of the general order of 10^{-13} cm. Yet again, it might be that space-time curvatures do not approach at all closely either of these values. But if not, this would require a more radical rethinking still of basic physical concepts, since the new physical theory would be needed at an even larger scale--a scale at which there has been as yet no hint from any

other direction that such a new theory is necessary for the performing of calculations.

One reason that such a new theory would be needed at some stage, however, is that the existing theory of space-time structure, namely classical general relativity, leads to the conclusion, via the "singularity theorems" (cf. [1],[2]), that under suitable qualitative circumstances singularities in space-time are inevitable. These qualitative circumstances would arise in (at least) two main situations, namely inside black holes and at the "big bang" origin of the universe. It is noteworthy that there is some observational evidence in support of both types of situation occurring in the actual universe. The observations do not directly imply the existence of singularities, but they do suggest that the aforementioned qualitative circumstances actually occur.

I shall not discuss the singularity theorems in detail here, as these have been amply described elsewhere. However, it is worth remarking on the main physical assumption involved (beyond the validity of general relativity itself) namely the energy condition (cf. [1],[2]). This states, in effect, that at each point the energy density component (to a comoving observer) of the energy momentum tensor plus the sum of the principal pressures is non-negative, and that the energy density plus each principal pressure is individually non-negative. If only the second of these assumptions is to hold, this is the weak energy condition. The weak energy condition is a consequence of the assumption that to every observer (not merely those comoving with the matter) the energy density is non-negative. The existence of space-time singularities under appropriate circumstances follows merely from the weak energy condition (cf. [3]), although only if a suitable Cauchy surface is assumed to exist.

So it is clear that the positivity of energy density is closely related, via Einstein's equations, to the existence of space-time singularities. It would seem that on the classical scale, such positivity is a necessary consequence of present-day physical ideas (such as that on some smaller scale quantum field theory is appropriate). However, as soon as the matter needs to be treated quantum mechanically on the scale under consideration, such positivity (or effective positivity) can no longer be assumed. Indeed, as the Hawking "black-hole explosion" phenomenon indicates (as we shall see later), there are reasons for believing that this positivity does not in fact hold at the quantum level. The question involved here is not one of quantum gravity, but of the nature of quantum fields on an effective classical curved space-time background. Even at this level there are fundamental questions of principle which remain unsolved.

Since the physics "at" a space-time singularity is not something that present-day theory can cope with, the precise mathematical discussion of singularities must concern the non-singular portions of the space-time only. This discussion may consist of an investigation of the way in which Einstein's equations

Space-time singularities; *Proceedings of the first Marcel Grossman meeting on general relativity.* Ed. R. Ruffini. ICTP Trieste North-Holland, Amsterdam, 1977

describe the evolution of space-time structure as a singularity is approached (cf. the work of the Soviet school [4],[5]) or it may be of a general character, in which the singular points are viewed, rigorously, as boundary points to the non-singular space-time manifold. It is this second approach that I shall adopt--although I shall be biassed here also, in that only one of the several different possible approaches ([6],[7],[8],[9],[10]) to attaching boundaries to space-times will be considered.

Suppose that the space-time M satisfies an appropriate causality condition (such as strong causality or, at least, future and past-distinguishability [10],[2]). Then each point p of M may be uniquely associated with the set of points $I^-(p)$ lying to the past of p in M (i.e. which can be connected to p by a future-timelike curve). If we know $I^-(p)$ then we know p and vice-versa. Such a set is called a PIP ('proper indecomposable past-set'). The PIP $I^-(p)$ may also be described as the set of points lying to the past of any time-like curve γ with future end-point at p. That is, $I^-(p) = I^-[\gamma]$. We say that γ <u>generates</u> the PIP $I^-[\gamma]$. Now the virtue of considering the set $I^-[\gamma]$ as a way of representing the point p is that it generalizes (see [10]). If γ is <u>any</u> timelike curve (not just one with a finite future end-point) then the set $I^-[\gamma]$ is called an IP ('indecomposable past-set'). If $I^-[\gamma]$ is not of the form $I^-(p)$ for $p \in M$, then $I^-[\gamma]$ is a TIP ('terminal indecomposable past-set'). Thus, timelike curves which go to infinity in the future generate TIP's. So do timelike curves which go to a singularity in the future. We may say that two timelike curves γ and γ' go into the future to the <u>same</u> point at infinity or to the <u>same</u> singular point if they generate the same TIP ($I^-[\gamma] = I^-[\gamma']$). With this criterion we may adjoin new "ideal" points to the space-time manifold, these being associated with the various TIP's. <u>Any</u> timelike curve now has a future "end-point". This may be an ordinary point of M (when $I^-[\gamma]$ is a PIP) or a future ideal point (when $I^-[\gamma]$ is a TIP). The future ideal points may be further classified as points at infinity or singular points according as the corresponding TIP's can, or cannot, be generated by timelike curves of infinite proper length into the future. (Alternative criteria can also be used.)

In an exactly similar way we can define sets $I^+[\gamma]$, called IF's ('indecomposable future-sets'), which are futures of timelike curves, these being either PIF's (if $I^+[\gamma] = I^+(p)$ for some $p \in M$) or TIF's (otherwise). The TIF's define the past ideal points for M. (Some ideal points may be considered to be both past and future ideal points at the same time, but I shall not bother with giving a definition for this here). These past ideal points are either points at infinity or singular points, according as the TIF's can or cannot be generated by timelike curves of infinite proper length into the past. In this way, any timelike curve in M acquires a past "end-point" which is either an actual point of M or an ideal point at infinity or a singular ideal point.

Space-time singularities; *Proceedings of the first Marcel Grossman meeting on general relativity.* Ed. R. Ruffini. ICTP Trieste North-Holland, Amsterdam, 1977

The ideal point structure of various space-times may be examined in detail. In the case when M is Minkowski space, for example, the different TIP's may be associated uniquely with the different points of future conformal infinity $I\!I^+$, together with the one point i^+ representing future timelike infinity. Similarly, the different TIF's are uniquely associated with the points of past conformal infinity $I\!I^-$ and i^-. In the case of the Friedmann models we obtain singular ideal points as well as (for some models) points at infinity. The singularities turn out to be three-dimensional sets. In fact, the various past ideal points, representing points of the initial big bang, correspond to the various particle horizons of the model (the sets $\partial I^+[\gamma]$). The singularity structure of the Schwarzschild solution may also be examined. For a black hole this singularity consists of future ideal points (TIP's), whereas for a white hole, it would consist of past ideal points (TIF's). In each case we again have a three-dimensional singular region. Also, as in the Friedmann models, this singular region is, in a well-defined sense, space-like. This can be seen from the fact that no singular TIP contains another singular TIP (for the future singularity) and no singular TIF contains another singular TIF (for the past singularity).

The question of the existence of timelike singularities in space-time models is closely related to what is known as the cosmic censorship hypothesis. A timelike singularity would (at least according to some definitions, cf. [11] be an example of a naked singularity--namely a singularity which lies to the future of some point on an observer's world line and to the past of another point on the same world line. (This is a local definition, which does not require reference to observers at infinity. It means, more technically, that a singular TIP is contained entirely in the past of some point of M, or, equivalently, that a singular TIF is contained entirely in the future of some point of M, cf. [11]. With this definition, the presence of a naked singularity implies the non-existence of a Cauchy hypersurface for the space-time. Other definitions are also used [2],[12]. Naked singularities are generally considered to be highly undesirable physically, since if they occur, an essential indeterminacy would be introduced into classical theory (over and above that already involved in quantum mechanics). The hypothesis of cosmic censorship asserts that naked singularities should not arise in any physically realistic classical situation.

There have been various attempts at obtaining a contradiction with the cosmic censorship hypothesis [12],[13], but so far none of these can be considered to have done so. Perhaps the failure of some of these attempts lends some small support to the validity of the hypothesis. On the other hand, I know of no argument which could be construed as a serious attempt at a proof of the hypothesis. So cosmic censorship remains a very open question--perhaps the most important unsolved question in general relativity theory.

If we exclude naked singularities, we are still left with the two

Space-time singularities; *Proceedings of the first Marcel Grossman meeting on general relativity.* Ed. R. Ruffini. ICTP Trieste North-Holland, Amsterdam, 1977

(spacelike?) types of singularity, namely those given by the past ideal points (big bang) and those given by the future ideal points (in black holes). Superficially, at least, there is a strong resemblance between these two types of singularity. It seems that apart from the difference in the scale involved, the past and future singularities might be essentially time-reverses of one another. On the other hand there is likely to be, in my view, at least one very important additional difference between these two types of singularity.

We have become accustomed to the fact that all the microscopic laws of physics, which have importance in governing the behaviour of matter as we know it (i.e. excluding K^0-decay), are symmetrical in time--whereas there is the gross statistical time asymmetry involved on a macroscopic scale associated with the phenomenon of entropy increase. The "blame" for the statistical asymmetry is to be placed on the boundary conditions. Thus the initial conditions of the universe must be such that the entropy is at a comparatively low value, while the final state would be one of much higher entropy. But to investigate the initial conditions of the universe we must examine the big bang singularity itself. If the universe is closed and recollapses to an all-embracing final singularity, then the entropy of the final singularity must be much greater than that of the initial singularity. (This is assuming that the statistical time-sense of the universe will not somehow reverse itself at the moment of maximum expansion. This possibility has been suggested by some authors, but I prefer to exclude it here. I find it very hard to reconcile with other aspects of our understanding of the physical world.) Accordingly, the nature of this final singularity ought to be very different from that of the initial singularity. The same would apply inside a black hole. To an observer falling into a black hole, the singularity inside the hole would appear all-embracing, like the singularity of a recollapsing closed universe (assuming naked singularities of the type referred to above are excluded) This all-embracing singularity would form his final boundary condition, so if his statistical time-sense is not to reverse as he falls into the hole (and I am excluding this possibility from consideration), this should be a high-entropy singularity, unlike that of the initial big bang.

To emphasize that the time-asymmetry ought to be present in the structure of the space-time singularity itself, not just in the nature of the matter in the neighbourhood of the singularity, it may be recalled that discussions of the nature of matter in the early stages of the universe tend to suppose the matter to be completely thermalized. Thus, as far as the matter itself is concerned, the entropy is as high as possible. If it were not for the degrees of freedom in the geometry of the space-time itself, there would be no room for this entropy to increase further. In some sense it must be that the entropy of the geometry starts off low--and ends up high.

Initially it is the expansion of the universe which allows room for the

entropy of the matter in it to increase. The entropy of the matter lags behind the value it would need to have for complete thermalization. And as the matter cools, local gravitational effects begin to play a role. Now the entropy can increase by local agglomerations of matter clumping together. In the way which is peculiar to gravity, higher and higher entropy values can be obtained as the matter clumps more and more. We obtain galaxies, normal stars, then white dwarfs or neutron stars. Finally, the greatest entropy of all is achieved when the matter collapses into black holes.

It appears that it is these gravitational degrees of freedom which are not used at all (or, at least, very little used) in the initial big bang. As far as can be inferred from observation (principally from the isotropy of the 2.7°K black-body radiation), the initial singularity was remarkably free of irregularity, the universe apparently resembling a Friedmann model to a high degree of approximation. This is quite unlike what seems reasonable to infer about the nature of a final singularity, whether it be that inside a black hole or one resulting from the recollapse of the universe as a whole. Any deviations from spherical symmetry in a collapsing body would be expected to get magnified as that body shrinks. (This is assuming, again, that the statistical time-sense does not reverse as the final singularity is approached!) The structure of the resulting singularity must be expected to be very irregular. Indeed, the analysis of Belinskii, Lifshitz and Khalatnikov [5] suggests a particularly complicated geometry for such singularities. The nature of the matter near the singularity, however, becomes quite unimportant. The geometry becomes more and more like that of a vacuum space-time, with the Weyl (conformal) part of the curvature ultimately completely dominating the Ricci part.

The following hypothesis thus suggests itself: the set of <u>past</u> ideal points has a particularly simple structure, the Weyl tensor vanishing in its immediate neighbourhood--on the other hand, at the <u>future</u> singular ideal points the Weyl tensor is likely to dominate. It seems that in some way the Weyl tensor gives a measure of the entropy in the space-time geometry. The initial curvature singularity would then be one with large Ricci tensor and vanishing Weyl tensor (zero entropy in the geometry); the final curvature singularity would have Weyl tensor much larger than Ricci tensor (large entropy in the geometry). These ideas are indeed somewhat imprecise, but it would seem worth while to pursue them further. The vanishing of Weyl tensor at the initial singularity, for example, would of itself seem to be effectively sufficient to ensure that the space-time closely resembles a Friedmann model, at least in its early stages. Of course, some way of introducing irregularities must also be allowed.

One consequence of the above hypothesis is that white holes are excluded. Even if exact spherical symmetry is assumed, the singularity inside the hole, though consisting of past ideal points, is quite different from that of the

Friedmann models. The Weyl tensor diverges to infinity rather than becoming zero. In any case, white holes must be thought of as anti-thermodynamic objects. If they are treated according to a physics subject to the normal statistical time directivity, then it seems that they are unstable [14],[15]. Furthermore, the area of the horizon of a white hole is subject to a law of surface area decrease. The Bekenstein [16] formula which states that the entropy of a black hole is proportional to the area of its horizon leads, if applied to a white hole, to the conclusion that the entropy of a white hole should decrease rather than increase. So it would seem satisfactory that a hypothesis governing the entropy of the universe as a whole should also directly exclude white holes.

At this point, mention should be made of some interesting ideas due to Hawking [17]. In accordance with the ideas of Bekenstein, a black hole ought also to possess a temperature proportional to its surface gravity (although this temperature would be absurdly low for any black hole which could arise by stellar collapse). Hawking predicts that owing to quantum mechanical effects, a black hole ought indeed to radiate as a black body of that temperature. If the temperature of the black hole exceeded that of its surroundings, it would radiate away some of its mass. This would mean that the area of its horizon should decrease, in contradiction with the classical area principle (which states that the area of the absolute event horizon of a black hole should be non-decreasing with time [2]). The occurrence of this phenomenon indicates that the weak energy principle is effectively being violated--because quantum rather than classical fields are involved. This decrease in surface area is now consistent with thermodynamics. The external temperature being lower than that of the black hole, the <u>total</u> entropy of black hole <u>plus</u> surroundings still increases. As the black hole gets smaller it also gets hotter. As it gets hotter it radiates faster, finally exploding with the emission of many baryons and anti-baryons. At its final disappearance, the black hole would momentarily become a naked singularity (not violating the classical cosmic censorship principle since this arises from an essentially quantum effect). If it disappears instantaneously, this naked singularity seems not to cause any real difficulty.

For a solar mass black hole this process would be absurdly slow, taking something of the order of 10^{55} times the present age of the universe for completion. Though the effect could only be of astronomical significance if very small black holes (of radius of the general order of 10^{-13} cm) had been created in the big bang--a possibility which seems at variance with the hypothesis suggested here--it nevertheless is of considerable theoretical interest.

Hawking argued [18] that the thermal equilibrium state of a system enclosed in a box with perfectly reflecting walls should (if the size of the box is suitable, given the total energy) be a black hole in stable equilibrium with black body radiation of the same temperature. He further suggests that being a

state of thermal equilibrium, such a configuration ought to be time-symmetric--one reason arising from an appeal to CPT invariance. Hence, a black hole ought to be physically indistinguishable from a white hole! Accordingly, the Hawking (quantum mechanical) emission process ought to be physically indistinguishable from the time-reverse of the normal (classical) accretion process of a black hole. He also argues that although the classical geometry of a black hole differs from that of a white hole (even outside their respective horizons) this does not invalidate their identification, since the geometry cannot be treated strictly classically (so it is claimed) in a situation such as this.

Such an identification would, however, be at variance with the hypothesis put forward here, since white holes are to be excluded but not black holes. I feel that the viewpoint I am presenting receives some support from consideration of the deviations from equilibrium which would occasionally occur with the configuration of a black hole in a perfectly reflecting box. By an initial fluctuation, the hole might emit too much radiation to remain in stable equilibrium with its surroundings. It would then heat up and finally disappear in a burst of baryons and anti-baryons. Eventually it would reform out of the background energy to achieve the original equilibrium maximum entropy state once more. However, it is exceedingly hard to believe that its probable mode of formation would be the time-reverse of this method of disappearance. Baryons and anti-baryons would have to be specially created out of the radiation and aimed simultaneously at a tiny point, as the initial step in the creation of the hole. Only subsequently would ordinary radiation be falling in.

So it seems that this equilibrium is not in fact time-symmetric. And if it is indeed true that white holes are to be excluded, we have the possibility to contemplate an essentially time-asymmetric physics--which might be going on virtually at a sub-microscopic level. If so, we would have small violations not only of T-invariance (recall K^0-decay!), but also of CPT-invariance. One might speculate on the closing of a logical loop. If T and CPT non-invariant interactions occur in microphysics, then though unimportant to phenomena now, they might nevertheless be important in the regimes neighbouring space-time singularities. The hypothesis of initially vanishing Weyl tensor could conceivably then have a basis in microscopic physical laws.

REFERENCES

[1] S. W. Hawking & R. Penrose, Proc. Roy. Soc. (Lond) A314 (1970) 529.

[2] S. W. Hawking & G. F. R. Ellis, The Large Scale Structure of Space-Time (Univ. Press., Cambridge, 1973).

[3] R. Penrose, Phys. Rev. Lett. 14 (1965) 57.

[4] V. A. Belinskii, I. M. Khalatnikov & E. M. Lifshitz, Adv. Phys. 19 (1970)

[5] V. A. Belinskii, I. M. Khalatnikov & E. M. Lifshitz, Soviet Phys. J.E.T.P. 62 (1972) 1606.

[6] H.-J. Seifert, Gen. Rel. & Grav. 1 (1971) 247.

[7] B. Schmidt, Gen. Rel. & Grav. 1 (1971) 269.

[8] R. Geroch, J. Math. Phys. 9 (1968) 450.

[9] B. Schmidt, Comm. Math. Phys. 36 (1974) 73.

[10] R. Geroch, E. H. Kronheimer & R. Penrose, Proc. Roy. Soc. (Lond) A327 (1972) 545.

[11] R. Penrose, in Confrontation of Cosmological Theories with Observational Data (Ed. M. S. Longair, I.A.U., 1974) 263.

[12] R. Penrose, Ann. N.Y. Acad. Sci. 224 (1973) 125.

[13] H. MUller zum Hagen, H.-J. Seifert & P. Yodzis, Comm. Math. Phys. 34 (1973) 135.

[14] Ya. Zel'dovich, in Gravitational Radiation and Gravitational Collapse (ed. C. DeWitt-Morette I.A.U. 1974) 82.

[15] D. Eardley, Death of White Holes in the Early Universe (Caltech Preprint 1974).

[16] J. D. Bekenstein, Phys. Rev. D.7 (1973) 2333, 9 (1974) 3292.

[17] S. W. Hawking, in Quantum Gravity (eds. C. J. Isham, R. Penrose & D. W. Sciama, Clarendon Press, Oxford, 1975) 219.

[18] S. W. Hawking (Caltech Preprint 1974).

87
THE TWISTED CAMEL (OR HOW TO GET THROUGH THE EYE OF A SINGULARITY). INCLUDES: MORE ON THE TWISTED CAMEL

Twistor Newsletter **4** (1977). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

Perhaps the first use of a blow-up in twistor contour integration.

138 The twisted camel (or how to get through the eye of a singularity). Includes: more on the twisted camel. *Twistor Newsletter* **4** (1977). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

§5.9 THE TWISTED CAMEL (or, how to get through the eye of a singularity)

by R. Penrose

I make, here, some remarks concerning the type of circumstance depicted in Figure A, for a function of one or more twistor variables, where two pole singularities pinch together, but at the same time the pinching is compensated by the presence of a zero, i.e. we are concerned with integrating an expression of the form A/BC over some contour, where some relevant common zero set of B and C is also a zero set of A.

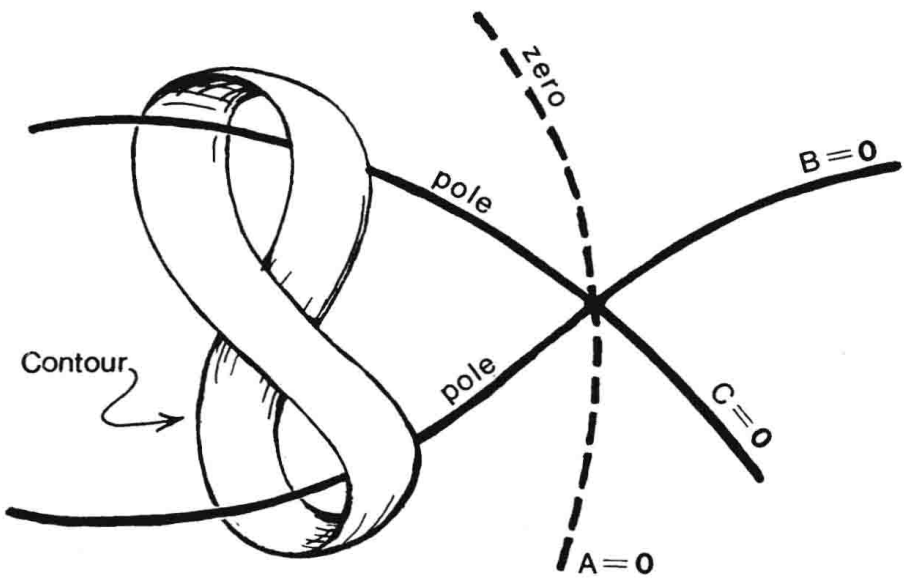


Figure A

The question of how to handle the contour integration when the contour approaches the point of pinching has some interest in a number of contexts. In particular, it has relevance to the integrals that A.P.H. considers in §5.8. In a sense, the contour can (at least in certain such circumstances)

be "pushed through" the point of pinching and give a finite result at all stages. I shall indicate *two* ways of so pushing our camel contour through the needle's eye: (i) take the needle away and put it somewhere else; and (ii) blow up the needle!

In method (i), the twistor function is regarded as part of a cochain; we are supposed not to be interested in the actual function itself but in some kind of cohomology group element that it defines. So the function can be replaced by another one in which (one hopes) the pinching now occurs somewhere else. A prototype of this type of behaviour arises when we try to represent certain generalizations of elementary states in terms of twistor functions, such "generalized elementary states" having arisen from a suggestion by M.F. Atiyah as to how best to apply R.S.W.'s Yang-Mills twistor construction (cf. §3.5). [See §5.8 for other uses of method (i) which could well be of considerable significance.] Such a generalized elementary state is obtained from some algebraic curve γ lying in PT^-, the point being that the field is to be singular *only* for complex space-time points which are described by lines in PT meeting γ (see Figure B).

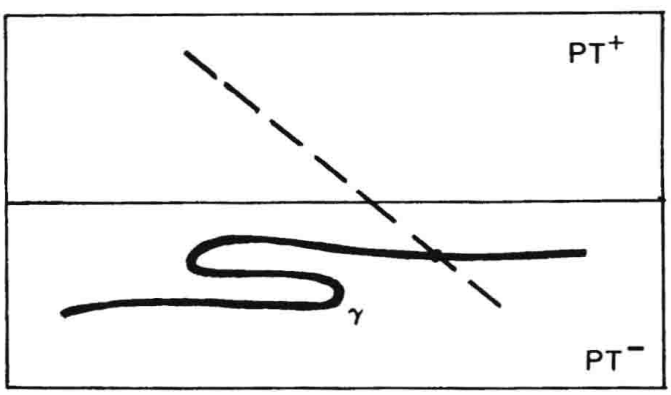


Figure B

 The twisted camel (or how to get through the eye of a singularity). Includes: more on the twisted camel. *Twistor Newsletter* **4** (1977). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

I shall consider only one example here, namely when γ is a *twisted cubic* (it had to happen!) and I shall not worry about arranging that γ actually lies in PT^- (which is easy enough to organize by means of a coordinate change). Choose twistor coordinates W, X, Y, Z and let γ be given by $W:X:Y:Z = 1:\theta:\theta^2:\theta^3$ with $\theta \in C\cup\{\infty\}$. Set $Q_1 = X^2 - WY$, $Q_2 = WZ - XY$, $Q_3 = Y^2 - XZ$. Then γ is the common intersection of the three quadrics $Q_i = 0$. Set $U_i = PT - \{Q_i = 0\}$. Then U_1, U_2, U_3 cover $PT - \gamma$. Set $f_{ij} = L_{ij}/Q_iQ_j$ where $L_{12} = W-X$, $L_{23} = Y-Z$, $L_{31} = X-Y$. Then the cocycle condition $f_{12} + f_{23} + f_{31} = 0$ is satisfied. Using f_{ij} as a twistor cocycle function, the field (a neutrino field) may be evaluated using a branched contour integral (see Figure C).

Figure C

But one always finds that the branched contour can be deformed until it is just *one loop* involving just *one* f_{ij}, for a generic space-time point. Pinching occurs only for exceptional space-time points - and then we simply *switch to a different* f_{ij}. Thus, we have an example of (i) above. For further discussion of massless fields based on a twisted cubic curve see §2.16.

To illustrate method (ii) I consider 2 dimensions first, say C^2, where the origin is the point of pinching. Take w,z as coordinates for C^2 and consider the integral

$$\oint_{\Gamma} f(w,z)\,dw\wedge dz \ ,$$

where

$$f(w,z) = \frac{A(w,z)}{B(w,z)C(w,z)} \ ,$$

the functions A, B, C having simple zeros on curves a, b, c through (0,0) with distinct tangents there. Consider the small sphere $w\bar{w} + z\bar{z} = \varepsilon^2$ ($\varepsilon > 0$). Holomorphic curves through (0,0) meet this S^3 in Clifford parallels (Hopf S^1's) (essentially) two of which are $b \cap S^3$ and $c \cap S^3$. The contour Γ can be a torus in S^3 avoiding and separating these S^1's, as depicted in Figure D.

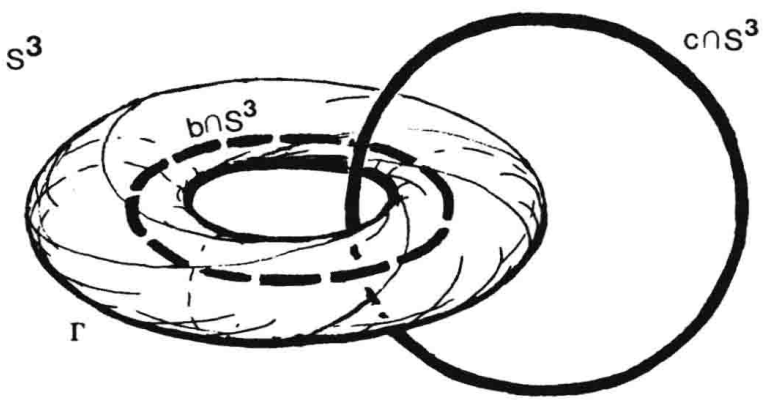


Figure D

If we now let $\varepsilon \to 0$, Γ shrinks down to the origin, which is singular: *pinching*!! So we *blow up* the origin, i.e. replace (0,0) by a whole S^2 (holomorphically!) in which *each* Clifford-Hopf S^1 converges down to a *different* point of S^2. The contour Γ can be fibred (if chosen suitably) by twisted Clifford-Hopf S^1's (cf. Robinson congruence pictures) each of which

collapses to a point as $\varepsilon \to 0$, but now b and c are *both avoided* even at $\varepsilon = 0$. So the contour slips off completely in the blown-up space! But we need to check that the form $\phi = f dw \wedge dz$ is non-singular on the new "origin" S^2 - and here's where A comes in. The blow-up can be achieved explicitly by mapping (w,z) to the point $(w^2 : wz : z^2 : w : z)$ of CP^4. Arrange coordinates so that $z = 0$ avoids a,b,c,Γ near (0,0), as can easily be done, and choose local coordinates $z = z^2/z$ and $q = w/z$ for the blown-up space. Then

$$\phi = \frac{A(zq,z)z\ dq \wedge dz}{B(zq,z)C(zq,z)},$$

which has a *finite* value on S^2 (which is defined by $z = 0$).

This shows that every contour integral of the above form must *vanish*, since the contour slips off through the blown-up origin.

A further illustration of method (ii) arises in connection with the spinor integral

$$\oint \frac{\pi.d\pi \wedge \lambda.d\lambda \wedge \mu.d\mu\,(\beta.\lambda)}{(\pi.\lambda)(\pi.\beta)(\lambda.\mu)^2}$$

examined by A.P.H. in §5.8. He observes that no contour exists, but nevertheless the integral can be given good meaning if $(\beta.\lambda)(\pi.\lambda)^{-1}(\pi.\beta)^{-1}$ is interpreted cohomologically. Another way of handling the "contour pinching" that arises here is to use a "blow-up", as suggested above. For this, first fix μ and consider the product space of (π,λ). This is a quadric Q in CP_3, and we have the set-up illustrated in Figure E:

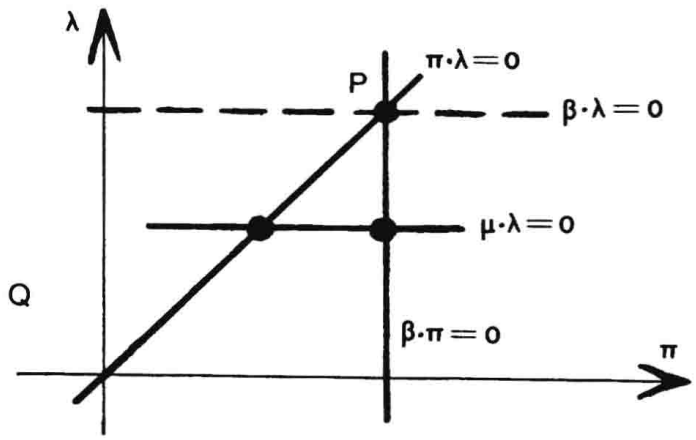


Figure E

The pinching that we wish to remove occurs where the poles $\pi.\lambda = 0$ and $\beta.\pi=0$ come together at the same place P as the zero $\beta.\lambda = 0$. The point P may be blown up into a line by sending *plane sections through* P, of Q, to *lines* in the new plane S. In the process, the two generators of Q through P (namely $\beta.\lambda = 0$ and $\beta.\pi = 0$) get "blown down" to points L and N in S. There remain but two singular lines in S, namely $\pi.\lambda= 0$ and $\mu.\lambda = 0$. (P is now non-singular because of the *numerator* $\beta.\lambda = 0$.) See Figure F.

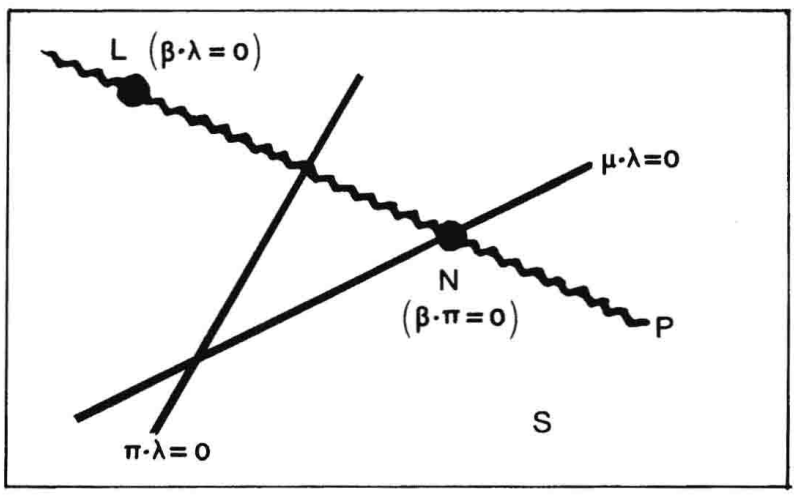


Figure F

Now let μ vary, so the line $\mu.\lambda = 0$ moves in our pictures. In the Q-picture, trouble occurs when this line passes across P, and any attempt at a contour gets pinched. But in the S-picture we can find an S^3 contour. This is Hopf-fibred by S^1's lying on lines through N, each S^1 looping once around N on the line. One such line is chosen for each position of $\mu.\lambda = 0$ - there is a continuous S^2's worth of them, having, say, antipodal intersections with $\pi.\lambda = 0$, to each intersection of $\mu.\lambda = 0$ with $\pi.\lambda = 0$. This whole S^3 may be identified with the boundary of a small 4-ball surrounding N in S.

88
REMARKS CONCERNING THE AMBIDEXTROUS PHOTON

Twistor Newsletter **5** (1977)

Remarks concerning the Ambidextrous Photon.

The information of a left-handed Maxwell field can be stored in a holomorphic line bundle over a suitable portion R of $\mathbb{PT}$ (RSW's twisted photon, TN1), the Euler operator Υ being well-defined on T and pointing along the fibres of T. Likewise, the information of a right-handed Maxwell field can be stored in a 1-manifold-fibration of T over R, but where now it is the holomorphic volume 4-form Δ ($= Dz$) which is well-defined on T instead of Υ (RP in TN3; cf. RSW in TN1). In TN3, LPH suggested putting the right-handed construction on top of the left-handed one, but it was not clear that this preserved any particular structure on T. In fact it does, namely the operator I shall call $\dot{\text{div}}$. This is a map from vector fields V, pointing along the fibration, to scalars, satisfying $\dot{\text{div}}(\lambda V) = \lambda \dot{\text{div}} V + V(\lambda)$. It differs from an ordinary div operation only in that it need not act on any vector field which does not point along the fibration. Clearly, if Δ exists, then $\dot{\text{div}}$ exists, defined by $\dot{\text{div}} V = \text{div}_\Delta V$, where div_Δ is the ordinary divergence based on Δ (i.e. use Δ to convert V into a 3-form $V \lrcorner \Delta$, take exterior derivative of $V \lrcorner \Delta$ and then divide out by Δ to obtain a scalar). Also, if Υ exists, then $\dot{\text{div}}$ exists, defined by $\Upsilon \dot{\text{div}} V = [\Upsilon, V] + 4V$.

Let us now suppose that the 4-dimensional space T, which is a fibration over R (all holomorphic) possesses just the structure div. Cover R by open sets $\{\mathcal{U}_i\}$ and lift these to give a covering $\{\mathcal{V}_i\}$ of T. Define a Δ_i and a Υ_i in each $\mathcal{V}_i$ which is compatible, in the above sense, with div. These compatibility requirements can be expressed as:

$$\operatorname{div} \Upsilon_i = 4 \quad \text{and} \quad \pounds_{\Upsilon_i} \Delta_i = 4\Delta_i \quad (\text{the latter being equivalent to } \operatorname{div}_{\Delta_i} \Upsilon_i = 4).$$

Now define $f_{ij} = \log(\Delta_j/\Delta_i)$. We have $\pounds_{\Upsilon_i} f_{ij} = 0$ and $\rho_{[i} f_{jk]} = 0$, so f_{ij} defines a 1-cocycle on R. This is the left-handed (hom. deg. 0) part of the photon. Whenever this cohomology class vanishes, Δ exists globally. Suppose, next, that Δ is global. Define $F_{ij} = (\Upsilon_j - \Upsilon_i) \rfloor \Delta$. We find $\pounds_{\Upsilon_i} F_{ij} = 0$ and $\rho_{[i} F_{jk]} = 0$, so F_{ij} defines a 1-cocycle 3-form on R. This is the right-handed (back-handed) (hom deg. -4) "part" of the photon.

In fact, this is exactly what we had before (in TN3 and TN1) although phrased somewhat differently, so that we don't even notice the $(\cdots)^{1/4}$ unpleasantness! (Incidentally, another way of disguising this is to write the $\hat{Z} = (1+4g)^{1/4} Z$ as $\hat{Z} = e^{g\mathcal{T}} Z$, where $g(Z)$ is hom. deg. -4 ... the power series *do* check!)

The trouble, of course, is the same as before: if there is a left-handed part (Δ not global) then there seems to be no way of extracting an $H^1(R, \mathcal{O}(-4))$ element to measure the non-globalness of Υ (without bringing in more structure, that is). This is the old problem — the right-handed part appears to be charged with respect to the left-handed part. To decouple the two parts, it seems that strict conformal invariance must be broken (e.g. $F_{ij} = (\Upsilon_j - \Upsilon_i) \rfloor \Delta_i$ or $F_{ij} = (\Upsilon_j - \Upsilon_i) \rfloor (\Delta_j + \Delta_i)$ won't do because the cocycle condition fails). Thus we must bring in I ("infinity twistor"). This works in principle, but is, as yet, inelegant.

~Roger Penrose

CHAPTER 88

89
MASSLESS FIELDS AND SHEAF COHOMOLOGY

Twistor Newsletter **5** (1977). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

In this article, a basic argument is presented showing that twistor functions really are representatives of holomorphic sheaf cohomology. This procedure is extended, filled out, generalized, and made respectable, by my colleagues in Chapter 129.

 Massless fields and sheaf cohomology: *Twistor Newsletter* **5** (1977). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

§2.6. MASSLESS FIELDS AND SHEAF COHOMOLOGY *by R. Penrose*

In §2.2 I indicated that twistor functions are really to be regarded as providing representative cocycles for sheaf cohomology, so the twistorial representation of a massless field of "helicity $\frac{n}{2}$" is an element of $H^1(X,\mathcal{O}(-n-2))$, where X is an (open) region in PT (say PT^+) swept out by the projective lines which correspond to the points of a suitable given (open) region Y in CM (say CM^+) where the field is defined. Woodhouse (§2.3) showed how the space-time field could be obtained from the $H^1(\ldots)$ element, when the $H^1(\ldots)$ element is defined by means of Dolbeault cohomology (i.e., "anti-holomorphic" forms (Morrow & Kodaira 1971)). A method due to Ward shows how to obtain the field from a Čech cohomology cocycle f_{ij} using an extension of Sparling's "splitting" formula (cf. §2.4). Ward's method essentially replaces the "branched contour integral" described in the first section. Basically, Ward's method is as follows:

Let $f_{ij}(Z^\alpha)$ be homogeneous of degree -n-2 (with $n \geq 0$) and defined on $U_i \cap U_j$ (with $\{U_i\}$ a covering of $X \subset PT$), $Z^\alpha = (\omega^A, \pi_A)$, $\omega^A = ix^{AA'}\pi_{A'}$, etc. Then $\pi_{A'} \ldots \pi_{E'} f_{ij}(Z^\alpha)$ (where π appears n+1 times) is homogeneous of degree -1 in π, so ("splitting") we get (ρ_i meaning "restriction," adopting a notational device due to Hughston, where square brackets denote skew-symmetrization):

$$\pi_{A'} \ldots \pi_{E'} f_{ij}(Z^\alpha) = \rho_{[i}\psi_{j]A'\ldots E'}(x^r, \pi_{R'}) ,$$

whence

$$\phi_{jA'\ldots D'}(x,\pi) := \pi^{E'}\psi_{jA'\ldots D'E'}$$

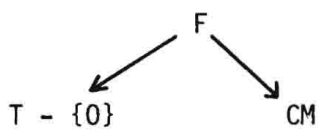
satisfies $\rho_{[i}\phi_{j]A'\ldots D'} = 0$. Thus, due to the sheaf property $\rho_{[i}F_{j]} = 0 \Rightarrow F_j = \rho_j G$, we have

$$\phi_{jA'\ldots D'}(x,\pi) = \rho_j \phi_{A'\ldots D'}$$

and, since $\phi_{A'\ldots D'}$ is homogeneous of degree 0 in π (and global in π), it must be a function of x only. Furthermore, $\nabla^{AA'}\phi_{A'B'\ldots D'} = 0$ readily follows. The case $n < 0$ is somewhat similar.

All this goes one way only: from twistor function to space-time field. But using the method of exact homology sequences we can readily extend this to obtain (implicitly) a version of the "inverse twistor function," which goes from space-time field to twistor function. (Compare also the Bramson-Sparling-Penrose method outlined in Penrose 1975.) The following was inspired to some extent also by Hughston (§2.5), cf. also Lerner (§2.7).

Let F be the (dual) primed spin-vector bundle (excluding the section of zero spin vectors) over CM. Thus, a point of F can be labelled $(x^a, \pi_{A'})$ (with $\pi_{A'} \neq 0$)-- except for points at infinity on CM--with x^a complex. F is also a bundle over $T - \{0\}$, the fibre over Z^α being the set of linear 2-spaces containing Z^α.



Consider the following exact sequence of sheaves over F:

$$0 \to \mathcal{T} \xrightarrow{\times \pi_{A'}\ldots\pi_{E'}} Z'_{n+1} \xrightarrow{\times \pi^{E'}} Z'_n \to 0. \qquad (1)$$

Here Z'_n stands for the sheaf of n-index symmetric spinor fields $\phi_{A'\ldots D'}(x^r, \pi_{Q'})$ which are holomorphic in x^r and $\pi_{Q'}$, and satisfy the

massless free-field equation $\nabla^{AA'}\phi_{A'B'\ldots D'} = 0$ ($\nabla_{AA'} = \dfrac{\partial}{\partial x^{AA'}}$, $\pi_{A'}$ being merely a passenger). The map $Z'_{n+1} \to Z'_n$ is simply

$\psi_{A'\ldots D'E'} \mapsto \phi_{A'\ldots D'} = \pi^{E'}\psi_{A'\ldots D'E'}$ and we must check that it is *onto*, i.e., that given $\phi\ldots$ we can always solve for $\psi\ldots$. One method is as follows: *First*, fix $\pi_{A'}$, and choose a primed spinor basis $o^{A'}$, $\iota^{A'}$ with $\pi^{A'} = o^{A'}$. We have $\psi_{0'\ldots 0'0'} = \phi_{0'\ldots 0'}$, $\psi_{1'0'\ldots 0'0'} = \phi_{1'0'\ldots 0'}$, $\ldots$, $\psi_{1'\ldots 1'0'} = \phi_{1'\ldots 1'}$, but $\psi_{1'\ldots 1'1'}$ is fixed only by the massless field equations

$$-\frac{\partial}{\partial x_{01'}}\,\psi_{1'\ldots 1'1'} = \frac{\partial}{\partial x_{00'}}\,\psi_{1'\ldots 1'0'} \left(= \frac{\partial}{\partial x_{00'}}\,\phi_{1'\ldots 1'}\right)$$

and

$$-\frac{\partial}{\partial x_{11'}}\,\psi_{1'\ldots 1'1'} = \frac{\partial}{\partial x_{10'}}\,\psi_{1'\ldots 1'0'} \left(= \frac{\partial}{\partial x_{10'}}\,\phi_{1'\ldots 1'}\right),$$

the integrability conditions for which come simply from the massless equations on $\phi_{A'\ldots D'}$. We can ensure a unique solution locally in CM by choosing (say) $\psi_{1'\ldots 1'1'} = 0$ on an initial 2-surface $x_{11'} = \text{constant}$, $x_{01'} = \text{constant}$. *Second*, allow $\pi^{A'}$ to vary locally in $\mathbb{C}^2 - \{0\}$; take $\iota^{A'}$ to depend holomorphically on $\pi^{A'}$ ($= o^{A'}$) (say $\iota^{A'} \propto \text{constant}$). The above prescription for finding $\psi_{A'\ldots D'E'}$ ensures that it is holomorphic in $\pi_{A'}$--though the dependence on $\pi_{A'}$ may be complicated owing to the fact that the functional dependence of $\phi\ldots$ on the components $x_{A\tilde{O}'}$ is π-dependent. (Incidentally, for simplicity, choose the *unprimed* basis *constant*.)

To see what T is we examine the kernel of the above map, i.e., find ψ's for which $\pi^{E'}\psi_{A'\ldots D'E'} = 0$ ($\pi^{E'} \neq 0$). We have (by standard lemmas) $\psi_{A'\ldots E'} = \pi_{A'}\ldots\pi_{E'}f$, and the massless equations on $\psi\ldots$ become

$\pi^{A'}\nabla_{AA'}f = 0$. Thus the dependence of f on x^a is only through the quantity $\omega^A = ix^{AA'}\pi_{A'}$ (i.e., f is constant on twistor planes in CM), so $f = f(\omega^A,\pi_{A'}) = f(Z^\alpha)$ is a *twistor function*. We see that $\mathcal{T}$ is the sheaf of twistor-type functions on F; equivalently, of holomorphic functions on T - {0}.

The short exact sequence (1) gives rise to the long exact homology sequence:

$$\ldots \to H^0(Q,Z'_{n+1}) \to H^0(Q,Z'_n) \to H^1(Q,\mathcal{T}) \to H^1(Q,Z'_{n+1}) \to \ldots \qquad (2)$$

Now, restrict Z'_n to be homogeneous of degree 0 in $\pi_{A'}$ (sheaf $Z'_n(0)$); correspondingly, Z'_{n+1} to be homogeneous of degree -1 in $\pi_{A'}$ (sheaf $Z_{n+1}(-1)$); so that $\mathcal{T}$ will be homogeneous of degree -n-2 in $\pi_{A'}$ (sheaf $\mathcal{T}(-n-2)$), i.e., of degree -n-2 in Z^α. Choose $Q \subset F$ to correspond to the region of F lying above $Y \subset CM$ (so, for each $(x^a,\pi_{A'}) \in Q$ we have $(x^a,\hat{\pi}_{A'}) \in Q$ whenever $\hat{\pi}_{A'} \in C^2 - \{0\}$). Then since H^0 means "global sections" and since Z'_{n+1} is homogeneous of degree -1 in $\pi_{A'}$ we have

$$H^0(Q,Z'_{n+1}(-1)) = 0.$$

To study $H^1(Q,Z'_{n+1})$ we can use, for example, a "resolution" of Z'_{n+1} (where dependence on $\pi_{A'}$ is irrelevant)

$$0 \to Z'_{n+1} \to F_{0,n+1} \to F_{1,n} \to F_{0,n-1} \to 0, \qquad (3)$$

this being exact, where $F_{p,q}$ is the sheaf of fields $\chi^{A\ldots C}_{D'\ldots F'}(x^r,\pi_{R'})$ with p symmetric unprimed indices and q symmetric primed ones (freely holomorphic), so $F_{p,q}$ is a coherent analytic sheaf, yielding $H^r(S,F_{p,q}) = 0$ whenever S is Stein and $r > 0$. The maps are:

$$\phi_{A'\ldots E'} \mapsto \chi_{A'\ldots E'} = \phi_{A'\ldots E'}{}^{\prime}$$

$$\chi_{A'B'\ldots E'} \mapsto \theta^{A}{}_{B'\ldots E'} = \nabla^{AA'}\chi_{A'B'\ldots E'}{}^{\prime}$$

$$\theta^{A}_{B'C'\ldots E'} \mapsto \eta_{C'\ldots E'} = \nabla^{B'}_{A}\theta^{A}{}_{B'C'\ldots E'}{}^{\prime}$$

and exactness is not hard to verify. If n=0, the sequence terminates one step sooner (and if n=-1 we also have a short exact sequence

$$0 \to Z'_0 \xrightarrow{i} F_{0,0} \xrightarrow{\Box} F_{0,0} \longrightarrow 0$$

where $\Box$ is the D'Alembertian). Note that (3) is conformally invariant, the conformal weights of $\phi\ldots$, $\chi\ldots$, $\theta\ldots$, $\eta\ldots$ being, respectively, -1, -1, -3, -4. The sequence (3) (for which $\pi_{A'}$ is irrelevant) is of interest in the study of massless fields quite independently of twistor theory. Here we just use it to derive the fact that

$$H^1(Q,Z'_{n+1}) = 0$$

whenever Y is Stein (as is the case if Y is the future- (or past-) tube in CM, or if Y is a suitable "thickening" of a portion of the real space-time M. (This follows because $H^1(Q,F_{0,n+1}(-1))$ then vanishes from the homogeneity degree -1 in $\pi_{A'}$, the fact that $H^1(CP^1,\mathcal{O}(-1)) = 0$ and suitable general theorems--for which thanks are due to R.O. Wells--and because $H^0(Q,F_{0,n+1}(-1)/Z'_n(-1))$ also vanishes--more obviously, because of -1 degree homogeneity.)

The upshot of this is that since the first and last terms in (2) both vanish (Y Stein), we have (for $n \geq 0$)

$$H^0(Q,Z'_n(0)) \cong H^1(Q,\mathcal{T}(-n-2))$$

which, because globality in π on the left-hand side implies constancy in π, almost establishes the required isomorphism between massless fields and H^1's of twistor functions. The remaining essential problem (the non-triviality and solution of which have been pointed out to me by M. Eastwood) is that all the sheaves refer to the space F, so far, and some subtleties are involved in projecting the cohomology groups down to T. The details of this and other matters are left to a proposed later paper to be written jointly with M. Eastwood and R.O. Wells, Jr. In fact the argument can be shown to work not only for the future tube (i.e., for $X = PT^+$) but also for suitable open regions in M. (Another subtlety has been glossed over in that the (x,π) description doesn't work for points at infinity. But this is unimportant because of conformal invariance and the fact that infinity can be transformed to somewhere safely inside the singular region of the field.)

To deal with the cases $n < 0$ we need a different exact sequence:

$$0 \to P_{m-2} \xrightarrow{i} T \xrightarrow{\hat{\pi}_A \cdots \hat{\pi}_C} Z_{m-1} \xrightarrow{\pi^{D'} \nabla_{DD'}} Z_m \longrightarrow 0. \qquad (4)$$

Here Z_m stands for the sheaf of *unprimed* massless holomorphic fields $\phi_{A\ldots D}(x^i, \pi_{R'})$ with $m = -n$ indices, and P_{m-2} stands for twistor functions which are polynomials in ω^A of degree at most m-2 for each fixed $\pi_{A'}$. The fact that $Z_{m-1} \to Z_m$ is onto, is basically an argument given in Penrose (1965); the kernel is this: twistor functions $\psi_{A\ldots C}(Z^\alpha)$, the massless field equations on which are $\pi^{E'} \hat{\pi}_{[D}\psi_{A]\ldots C} = 0$, yielding $\psi_{A\ldots C} = \hat{\pi}_A \ldots \hat{\pi}_C f(Z)$ as required. The short exact sequence

$$0 \to T/P_{m-2} \to Z_{m-1} \to Z_m \to 0$$

 Massless fields and sheaf cohomology: *Twistor Newsletter* **5** (1977). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

provides the long exact sequence

$$H^0(Q,Z_{m-1}) \to H^0(Q,Z_m) \to H^1(Q,T/P_{m-2}) \to H^1(Q,Z_{m-1}). \tag{5}$$

Choosing homogeneities -1, 0, m-2, -1, respectively, and using the same argument as above, we get (Y Stein):

$$H^0(Q,Z_m(0)) \cong H^1(Q,T(m-2)/P_{m-2}(m-2)). \tag{6}$$

To deal with the extra complication of P_{m-2} we use $0 \to P_{m-2} \to T \to T/P_{m-2} \to 0$ to derive the exact sequence

$$H^1(Q,P_k(k)) \to H^1(Q,T(k)) \to H^1(Q,T(k)/P_k(k)) \to H^2(Q,P_k(k)) \tag{7}$$

(where $k = m-2$). Now the sheaf $P_k(k)$ vanishes unless $k \geq 0$ (i.e., $m \geq 2$) and, using a suitable Kunneth formula, we derive the fact (since homogeneity in $\pi_{A'}$ for each x is non-negative) that we are concerned only with polynomial behaviour in $\pi_{A'}$ for the two end terms in (7). Consequently these terms refer, *in effect*, to the "ordinary" cohomology of Q but with coefficients which are twistor polynomials $Z^\alpha \ldots Z^\gamma P_{\alpha\ldots\gamma}$, homogeneous of degree k, i.e., equivalently, symmetric n-twistors $P_{\alpha\ldots\gamma}$. If Y describes a space-time region surrounding a "source tube," then we have $\frac{1}{6}(k+1)(k+2)(k+3) = \frac{1}{6}m(m^2-1)$ independent complex "charges," corresponding to the various independent components of $P_{\alpha\ldots\gamma}$, all of which have to *vanish* if the $H^1(Q,T)$ description is to work. This corresponds to the final map in (7) mapping to the zero element. (Note that m=2 for the anti-self-dual Maxwell case, which is consistent with a statement made in §2.2 about charge integrality). If Y is the forward tube CM^+, then the first and last terms of (7) both vanish and the required isomorphism follows (but various details remain to be worked out).

Alternative approaches to the cases $n < 0$ can be given in which a potential rather than the field is used. This avoids having to bring in P_k and $P_{\alpha\ldots\gamma}$. (Work by Ward, Hughston and Eastwood.)

Further work is in progress.

Thanks are due to M.F. Atiyah, M. Eastwood, R.O. Wells, Jr. and R.S. Ward.

References

Morrow, J. & Kodaira, K. 1971 *Complex Manifolds*. Holt, Rinehart & Winston.

Penrose, R. 1965 Proc. Roy. Soc. A284, 159-203.

Penrose, R. 1975 *Twistor Theory, its Aims and Achievements*. In: *Quantum Gravity*, eds. C.J. Isham, R. Penrose & D.W. Sciama, pp. 268-407. Clarendon: Oxford.

90
NOTE ON THE N-TWISTOR INTERNAL SYMMETRY GROUP

Twistor Newsletter **6** (1977). With G. A. J. Sparling. Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

See Chapter 66. Here the general case of the n-twistor internal symmetry group is considered. Later, George Sparling showed that this description gives the entire group, not just the identity-connected component.

 Note on the n-twistor internal symmetry group. *Twistor Newsletter* **6** (1977). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

§4.7 A NOTE ON THE n-TWISTOR INTERNAL SYMMETRY GROUP *by R. Penrose and G.A.J. Sparling*

Consider a system of n twistors Z_i^α (i = 1,...,n) with complex conjugates $\bar{Z}_\alpha^i$. The *kinematic twistor* (i.e. the "angular momentum twistor") $A^{\alpha\beta}$ for this system is defined by

$$A^{\alpha\beta} = 2Z_i^{(\alpha}I^{\beta)\gamma}\bar{Z}_\gamma^i$$

The *n-twistor internal symmetry group* (or, simply, the *n-twistor group*) is defined to be the group of complex linear transformations of Z_i^α and $\bar{Z}_\alpha^i$ that preserve $A^{\alpha\beta}$ and the relation of complex conjugacy between Z_i^α and $\bar{Z}_\alpha^i$.

For some time it has been known that the transformations

$$T: \quad \begin{pmatrix} Z_i^\alpha \\ \bar{Z}_\alpha^i \end{pmatrix} \longrightarrow \begin{pmatrix} U_i^j\delta_\beta^\alpha & \Lambda_{ik}\bar{U}_j^k I^{\alpha\beta} \\ \bar{\Lambda}^{ik}U_k^j I_{\alpha\beta} & \bar{U}_j^i\delta_\alpha^\beta \end{pmatrix} \begin{pmatrix} Z_j^\beta \\ \bar{Z}_\beta^j \end{pmatrix} ,$$

with U_i^j unitary and Λ_{ij} complex skew-symmetric, belong to the n-twistor group, and it has been suspected that they constitute the entire n-twistor group. We shall outline here a proof that, indeed, they constitute the entire identity-connected component of the n-twistor group.

The multiplication law for the group can be expressed as

$$(U_i^j, \Lambda_{k\ell}) \circ (U'^j_i, \Lambda'_{k\ell}) = (U_i^m\, U'^j_m,\ \Lambda_{k\ell} + U_k^m\, U_\ell^n\, \Lambda'_{mn}),$$

an "inhomogeneous U(n)". This group acts as

$$Z_i^\alpha \to U_i^j Z_j^\alpha + \Lambda_{ik}\bar{U}_j^k I^{\alpha\beta}\bar{Z}_\beta^j ,$$

together with the corresponding complex conjugate transformation. Our argument consists of showing that the infinitesimals of the n-twistor group

have the form of infinitesimal T's:

$$Z_i^\alpha \to Z_i^\alpha + i\varepsilon\, H_i^j Z_j^\alpha + \varepsilon L_{ij} I^{\alpha\beta} \bar{Z}_\beta^j \quad ,$$

with H_i^j Hermitian and L_{ij} skew-symmetric. Certainly the infinitesimal elements of the n-twistor group have the form

$$Z_i^\alpha \to Z_i^\alpha + \varepsilon(R_{i\beta}^{j\alpha} Z_j^\beta + S_{ij}^{\alpha\beta} \bar{Z}_\beta^j) \quad ,$$

and preservation of the kinematical twistor leads to:

$$I^{\alpha(\beta} S_{ij}^{\gamma)\delta} + I^{\delta(\beta} S_{ji}^{\gamma)\alpha} = 0 \quad , \tag{1}$$

$$R_{k\rho}^{i(\alpha} I^{\beta)\sigma} + \delta_\rho^{(\alpha} I^{\beta)\gamma} \bar{R}_{k\gamma}^{i\sigma} = 0 \; . \tag{2}$$

By symmetrizing (1) over $\alpha\beta\gamma$ we deduce that $S_{ij}^{(\alpha\beta)} = 0$. Then contracting (1) with $I_{\alpha\rho}$ we obtain expressions which, by the use of simple lemmas, give

$$S_{ij}^{\alpha\beta} = L_{ij}\, I^{\alpha\beta}$$

for some skew-symmetric L_{ij}. By applying somewhat similar operations to (2) and using rather more involved lemmas we finally derive

$$R_{k\beta}^{j\alpha} = i\delta_\beta^\alpha H_k^j$$

for some Hermitian H_k^j. □

The quantized twistor operators obtained from these infinitesimal generators define the algebra which gives the classification of *n-twistor particles* (quantization: $\bar{Z}_\alpha^i \to -\partial/\partial Z_i^\alpha$). For n = 3 this algebra is the same as that of Z.P.'s inhomogeneous SU(3). Details will be described elsewhere. For further discussion (and references) see §4.8.

91
ASPECTS OF GENERAL RELATIVITY

Physics and contemporary needs, Vol. 1 (1977), 383–426. Ed. Riazuddin. Plenum, New York, 1977

This series of talks, given in Nathiagali, Pakistan, in August 1976, appears to contain one of the very first mentions of the 'Weyl curvature hypothesis', this characterising the low-entropy nature of the big-bang singularity as having vanishing (or at least highly constrained) Weyl curvature, as opposed to the expected unconstrained wildly diverging Weyl curvature of the singularities arising from gravitational collapse. See also Chapters 86 and 96. Something odd seems to have happened to some of my (evidently redrawn) diagrams in this article; most notably, Figure 20 on p. [414] appears to be upside-down.

164 Aspects of general relativity. *Physics and contemporary needs*, Vol. 1 (1977), 383–426.
Ed. Riazuddin

ASPECTS OF GENERAL RELATIVITY*

Roger Penrose

Rouse Ball Professor of Mathematics

Oxford University, Oxford, England

1. The Present Observational Status of General Relativity Theory

While the effects of Einstein's general relativity are not observationally significant for the every day physics of terestial gravity, nor (in almost all cases) for the motion of planets about the sun, nor (as far as we know) for stars within galaxies, nor for galaxies about one another, the theory does give some observed (or observable) corrections to Newtonian theory in suitable circumstances. In addition to the 'classical' observed perihelion advance of Mercury (and, to a lesser extent of other planets as well) and the observed bending of light by the sun's gravitational field, general relativity contributes significantly to the structure of neutron stars, to the stability of stellar models and to dynamical effects in close binary systems owing to energy loss in gravitational radiation. There are also effects which, though small in general relativity, are absent altogether in standard Newtonian theory, such as the slowing down of clocks in a gravitational potential, the time delay in light signals passing near the limb of the sun (both observed effects) and the very existence of energy-carrying gravitational waves (whose observational status is more dubious). Moreover, there are situations in which general relativity, effectively, provides a gross new effect, such as in the structure of black holes (an example of which apparently resides in the X-ray source Cygnus X-1), the global structure of the Universe (i.e. cosmology), the existence of unavoidable space-time singularities in black-holes

---

* Notes taken by A.Qadir, Quaid-i-Azam University, Islamabad.

and at the beginning of the **Universe** (the 'big bang' of the cosmological models). Also, there are new effects to be expected when quantum theory and general relativity are brought to bear upon one another. When gravitational fields are strong enough (which would be the case only in the neighbourhood of singularities), particle creation is to be expected. One example of this occurs with the Hawking process, according to which a black-hole emits thermal radiation of a temperature inversely proportional to its mass. However, this could be observationally significant only if very tiny black holes were created in the 'big bang'.

Though the observational data supporting general relativity are still rather scanty, they are sufficient, when combined with certain important theoretical requirements, to rule out (or at least to render exceedingly implausible) virtually all its rivals. The most serious alternative is the Brans-Dicke theory, but even this theory is compatible with the observations only if its free parameter is chosen to have such a large value that the theory becomes observationally undistinguishable **from Einstein's. As of now** there are no unambiguous observations inconsistent with the predictions of general relativity.

2. Spatial Geometry of Cosmological Models

It is often very helpful to employ geometrical arguments in relativity, either in place of, or else as a supplement to analytical ones. For example the spatial geometry of the standard Friedmann-Robertson-Walker models of the **Universe** may be studied geometrically. The spatial metric is

$$d\sigma^2 = \frac{dx^2 + dy^2 + dz^2}{1 + \frac{1}{4} K(x^2+y^2+z^2)} \qquad (1)$$

where K = +1, 0, or -1. Let us consider the three cases separately.

(i) K = +1. Spatially spherical Universe.

We can try to visualize a three dimensional sphere, S^3, (i.e. a spherical 3-surface in a Euclidean 4-space) by considering the projection of the sphere from the north pole into a flat three dimensional space, as shown in Figure 1 (with two dimensions suppressed). The ratio of the metrics for the Euclidean three dimensional space, E^3, to that for the three dimensional sphere (since the plane element at P and Q make equal angles with NQ) is

$$\frac{d\sigma(E^3)}{d\sigma(S)} = \frac{NQ}{NP} = \frac{NP + PQ}{NP} = 1 + \frac{r^2}{4} \qquad (2)$$

 Aspects of general relativity. *Physics and contemporary needs*, Vol. 1 (1977), 383–426.
Ed. Riazuddin

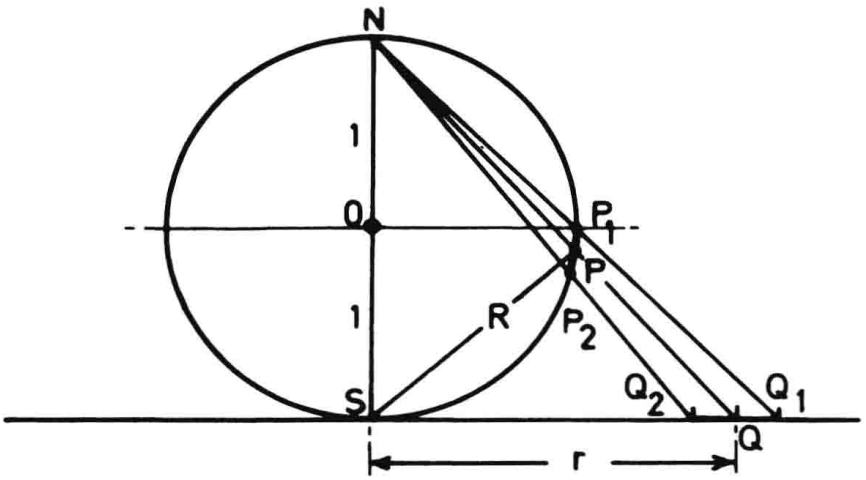


Figure 1. Stereographic projection of S^3 (The arc P_1 P P_2 is mapped into the line Q_1 Q Q_2).

(ii) K = 0.Spatially Euclidean Universe

This is the ordinary flat three dimensional Euclidean space, E^3.

(iii) K = -1.Spatially Hyperbolic Lobachevski Universe.

This may well turn out to be the type of the actual Universe. It is a (Sacceri-Gauss-Bolyai)Lobachevski space (L^3). Suppressing one dimension, we may think of it as being represented by one sheet of a hyperboloid (the 'mass hyperboloid') in Minkowski space,(the space-time of special relativity) namely the set of points of unit Minkowski-distance from the origin and lying to the future of the origin. We obtain a projective or a conformal representation of L^3 by projecting from the points (0,0,0,0) and (-1,0,0,0), respectively, into the plane T = 1, as shown in Figures 2, 3a and 3b. The metric of the Lobachevski space, L^3, can be represented as that induced by the Minkowski space-time metric

$$ds^2 = dT^2 - dX^2 - dY^2 - dZ^2 = -d\sigma^2 \qquad (3)$$

on the hyperboloid sheet

$$T^2 - X^2 - Y^2 - Z^2 = 0 \qquad (T > 0) \qquad (4)$$

(This being a space-like hypersurface, $d\sigma^2$ is a positive definite 3-metric. The variable T may be eliminated by using Eq.(4), if desired, and $d\sigma^2$ expressed solely in terms of X, Y, Z). The projections into T = 1 from (0,0,0,0) and from (-1,0,0,0) represent L^3

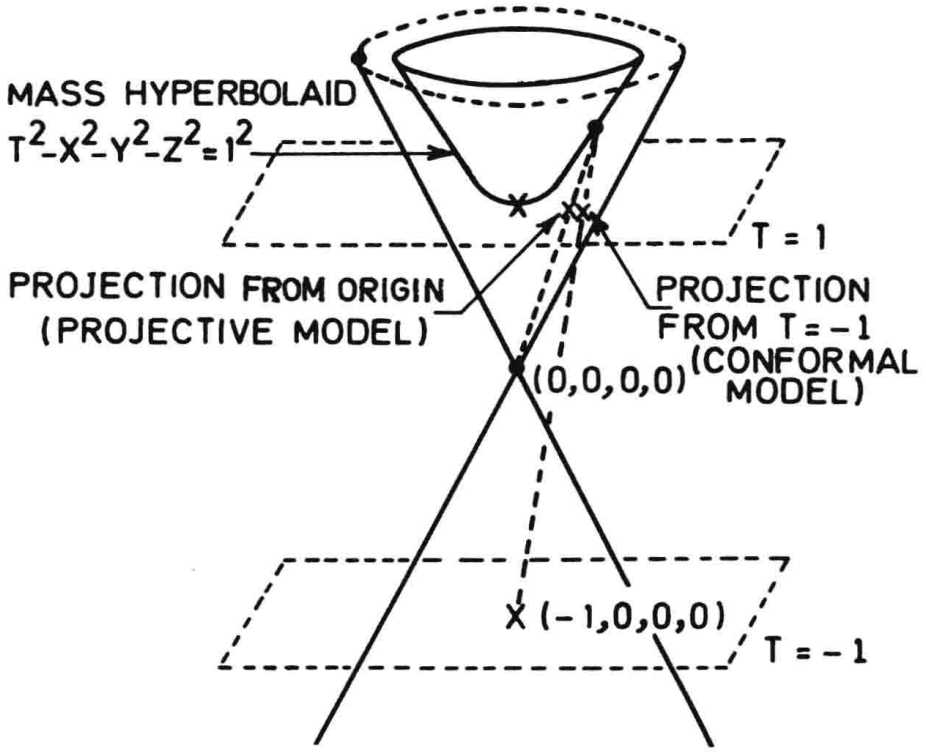


Figure 2. The Lobachevski space L^3.

as the region $X^2 + Y^2 + Z^2 < 1$ and < 4 respectively. Let us scale the second case down by a linear factor 2 (or, equivalently, project into T = 0 rather than T = 1); then in both representations L^3 is described by the unit ball in E^3. Let p be the distance of a point from the centre in the first (projective) representation and C the distance of the corresponding point in the second (conformal) representation. Then we can readily verify that

$$p = \frac{2C^2}{1 - C^2} \qquad (5)$$

the two points being in the same straight line through the origin. The merit of the projective representation is that geodesics are described by straight lines. In the conformal representation, geodesics are described by circles which intersect the boundary orthogonally. The special merit of the conformal representation is that angles (and small shapes) are depicted without distortion. The distance between two points B and C (see Figures 3a and 3b) in <u>each</u> model is given by

 Aspects of general relativity. *Physics and contemporary needs*, Vol. 1 (1977), 383–426. Ed. Riazuddin

$$d = \tfrac{1}{2} \log \left(\frac{AC \cdot BD}{AB \cdot CD} \right) \tag{6}$$

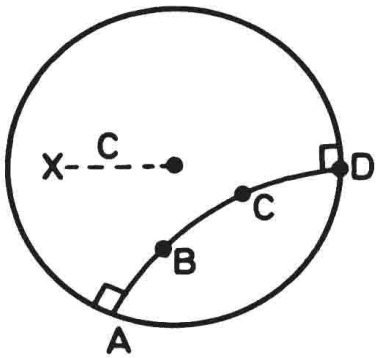


Figure 3a. Conformal model.

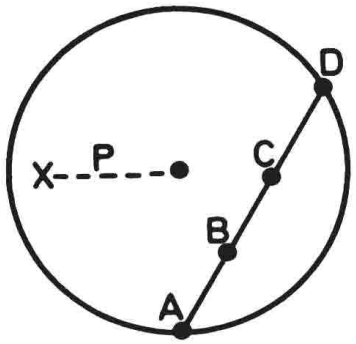


Figure 3b. Projective model.

3. Space-time Geometry

We now come to the geometrical description of space-time. In special relativity we use Minkowski (i.e. a flat) space-time. This means that there are suitable (i.e. Cartesian) co-ordinates for which the metric is given by Eq.(3). (We shall be using gravitational units $c = 1 = G$ or 1 sec. $= 3 \times 10^{10}$ cm $= 4 \times 10^{38}$ gm.) The cone in Figure 2 is called the null cone of the origin. The paths of all particles passing through the origin must be contained inside the cone, i.e.

$$T^2 - X^2 - Y^2 - Z^2 \geq 0. \tag{7}$$

Similarly there is a null cone at any other point (T_o, X_o, Y_o, Z_o) constraining the motions of particles through that point to satisfy the inequality

$$(T - T_o)^2 - (X - X_o)^2 - (Y - Y_o)^2 - (Z - Z_o)^2 \geq 0. \tag{8}$$

Normally this is stated in infinitessimal terms; i.e. the metric expression in Eq.(3) is always greater than or equal to zero for real particles, and so the tangent vector to the path of a real particle must have positive or zero 'length squared'. In general co-ordinates the Minkowski metric would take on the more general form

$$ds^2 = g_{ij} \, dx^i \, dx^j \tag{9}$$

(where we have used the Einstein convention that repeated indices must be summed over).

However, not every metric of the form given in Eq.(9) (where the g_{ij} are smooth functions of x^0, x^1, x^2, x^3) represents Minkowski space. The condition that it does so is simply that some co-ordinates exist for which the original form (Eq.(3)) holds. Nevertheless (assuming that the matrix g_{ij} is everywhere non-singular and with the appropriate + - - - signature) the metric can always be transformed to the form given in Eq.(3) at any one point; that is to say, special relativity holds locally and hence a null cone always exists. As before, particle world-lines are constrained to lie within (or on) the null cones, i.e. the expression in Eq.(6) must always be non-negative, this being the condition that the local light velocity is not exceeded. In special relativity we can always choose our co-ordinate grid so that the null cones are depicted with semi-angle 45° and axis along the T-coordinate lines (i.e. the form given in Eq.(5) holds globally). In general relativity this is not generally possible and the situation resembles that shown in Figure 4.The

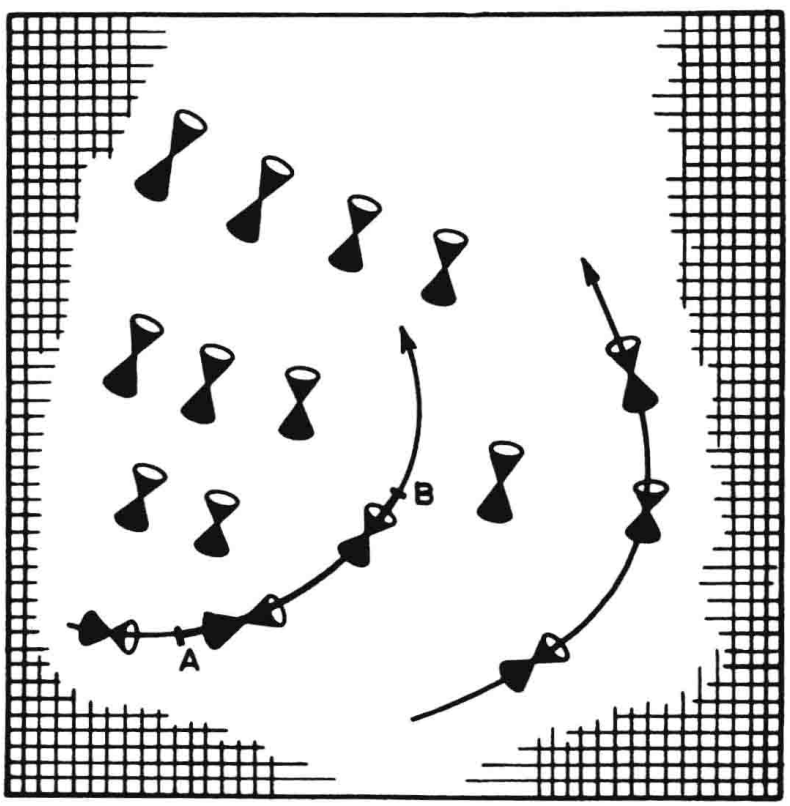


Figure 4. Geometrical description of space-time.

null cones may appear as squashed or stretched in one direction or another, or as tilted with respect to the co-ordinate grid. But it should be emphasized that this has no local significance. The co-ordinate grid is introduced as a convenience only and may always be replaced by another one in which the squashing ant tipping looks different. Nevertheless, such space-time pictures are invaluable for the discussion of the global structure of space-time.

We define the proper time (i.e. the time measured by an atomic clock attatched to an observer) to be the integral along the path of the observer of the metric (given in Eq.(9)). Thus the proper time from an event (a point in space-time) A to an event B is:

$$\text{Proper time along the world-line from A to B} = \int_A^B ds \qquad (10)$$

This description has been found to be valid up to accelerations of 10^{16} gm experimentally.

Einstein pointed out that the equivalence between the gravitational mass m_{grav} and the inertial mass m_{iner} given by

$$F = \frac{G M_{grav} \; m_{grav}}{r^2} \qquad (11)$$

$$F = m_{inert} \; a \qquad (12)$$

respectively, could not be an accident to be ignored, but must have a deep significance. (It should be noted that the corresponding property for electromagnetism is not true, i.e. $m_{inert} \neq e$.) He therefore postulated the principle of equivalence, which may be expressed by the statement that the gravitational force can be locally eliminated, i.e. there always exists a local frame (called the freely falling frame) in which there is no gravitational force. The frame was, in effect, shown to exist by the researches of Galilleo (as, according to tradition, he is supposed to have dropped two stones of different masses from the leaning tower of Pisa, each appears to be at rest in the frame of rest of the other) and is dramatically seen to exist by an astronaut out on his space-walk.

4. The Concept of a Geodesic

A geodesic is the curved space generalization of a straight line. In a curved space with a positive definite Riemannian metric a geodesic locally minimizes the arc length integral. With the hyperbolic + - - - signature of the metric of space-time, the arc-length is merely stationary for a space-like geodesic. But a time-like geodesic is characterized by the fact that the proper-time integral is locally maximized along the curve. This applies to the free motions of all particles. For the particular case of Minkowski

space it is readily seen that a particle at rest will give maximum proper-time for all allowable paths connecting two points on the time-axis (given by putting dx = dy = dz = 0 in Eq.(3)). This is the mathematical expression of the so called clock paradox: the proper-time experienced by an observer stationary in a Minkowski frame is always larger than that experienced by any other accelerating (i.e. non-geodesic) observer who starts co-incident with the first observer and finally returns. In general relativity the situation is basically the same locally, but new features can arise in the large. The question will be returned to later.

The differential equation for a geodesic can be given in terms of the operation of co-variant derivative, ∇_a, (which is also denoted by ";a", "|a", "||a"). This operator acts on tensors to produce new tensors. In particular, when acting on a scalar function, ϕ, it simply yields the ordinary gradient

$$\nabla_a \phi = \frac{\partial \phi}{\partial x^a} \quad (=\phi_{,a}) \tag{13}$$

but acting on a vector it requires an additional term

$$\nabla_a V^b = \frac{\partial V^b}{\partial x^a} + \Gamma^b_{ac} V^c \tag{14}$$

and generally, acting on a tensor gives

$$\begin{aligned} \nabla_a T^{b\cdots d}{}_{e\cdots h} = {} & \frac{\partial T^{b\cdots d}{}_{e\cdots h}}{\partial x^a} + \\ & + \Gamma^b_{ay} T^{y\cdots d}{}_{e\cdots h} + \cdots \\ & + \Gamma^d_{ay} T^{b\cdots y}{}_{e\cdots h} - \\ & - \Gamma^y_{ae} T^{b\cdots d}{}_{y\cdots h} - \cdots \\ & - \Gamma^y_{ah} T^{b\cdots d}{}_{e\cdots y} \end{aligned} \tag{15}$$

where Γ^c_{ab} is the Christoffel symbol

$$\Gamma^c_{ab} = \frac{1}{2} g^{cy} (g_{ay,b} + g_{by,a} - g_{ab,y}) \tag{16}$$

g^{ab} being the inverse of g_{ab} (i.e. $g^{ay} g_{yb} = \delta^a_b$).

If we consider a curve parameterized by u, its tangent vector at a point $P(x^a)$ will be given by

 Aspects of general relativity. *Physics and contemporary needs*, Vol. 1 (1977), 383–426. Ed. Riazuddin

$$t^a = \frac{dx^a}{du} . \tag{17}$$

Then the equation of a geodesic (with affine parameter u) is

$$t^a \nabla_a t^b = 0 \tag{18}$$

i.e. the tangent vector is propagated along the curve parallel to itself. Writing this out, we obtain

$$\frac{d^2x^a}{du^2} + \Gamma^a_{bc} \frac{dx^b}{du} \cdot \frac{dx^c}{du} = 0 \tag{19}$$

which is the geodesic equation.

The principle of equivalence can be rephrased as the statement that free-motion under gravity (for a small spherical particle) is always along a space-time geodesic. Any other motion is always accelerated in an absolute sense, and it requires some non-gravitational force to be present in order for a particle to follow such a path.

5. The Non-Minkowski Nature of ds^2

Consider a photon of energy E moving up in the Earth's gravitational field. Clearly it should lose energy (due to the effective mass of the photon E/c^2) and at a later time (being at a higher gravitational potential) it would have energy $E' < E$. Now,

$$E = h\nu \quad , \quad E' = h\nu' . \tag{20}$$

Thus $\nu' < \nu$. But the photons would be emitted continuously from the lower potentials to be viewed at the higher potential. The wave fronts must keep pace with one another. Thus, regarding h as simply a conversion factor between units of energy and time, we see that the time scale at the higher potential must be such that the photons' oscillations are measured to have a lower frequency, that is to say, a time interval measured at the higher potential is greater than the corresponding time-interval measured at the lower potential. But the measurements of time intervals between events determine the metric, so the metric can not be that of Minkowski space, showing that we have to use a curved space-time in our description.

This argument was presented graphically by Bondi by means of a thought experiment in which excited atoms are induced to emit photons while at the bottom of a continuous chain, return to the top as unexcited atoms, absorb the photons while at the top and come

down again as excited atoms (being heavier by the principle of equivalence, as $E = mc^2$ (see Figure 5)). This mechanism would provide a perpetual motion machine - which is impossible - if the same photons

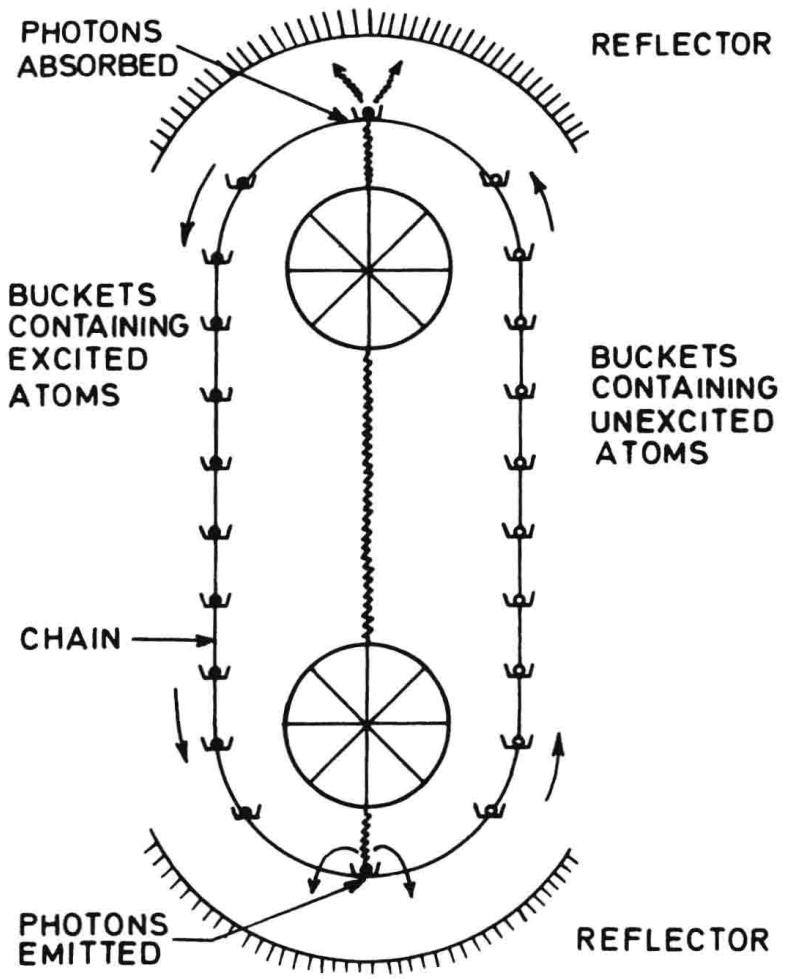


Figure 5. The Bondi buckets.

which are emitted at the bottom could also be reabsorbed at the top. The fact that they can not be reabsorbed is a manifestation of the above gravitational red-shift. This effect has been experimentally checked by Pound and Rebka (1960) and by Pound and Snider (1964) by sending light (γ-radiation) up a 70 feet tower and measuring the frequency difference by using the Mossbauer effect. The Pound-Rebka-Snider measurement may be taken as direct experimental verification of the non-flat nature of the space-time metric.

Aspects of general relativity. *Physics and contemporary needs*, Vol. 1 (1977), 383–426. Ed. Riazuddin

6. The Curvature Tensor

The principle of equivalence must not be taken to imply that the difference between gravitation and acceleration can never be experimentally measured. Gravitational effects are not locally observable (i.e. by studying a single geodesic), but can be observed by considering a family of geodesics, e.g. by examining tidal forces. The Newtonian picture of how tidal forces arise is shown in Figure 6. But in general relativity we may view the gravitational tidal effect as a direct physical manifestation of the curvature of space-time. It can be computed from the curvature tensor $R^a{}_{bcd}$ which is given by

$$(\nabla_c \nabla_d - \nabla_d \nabla_c) V^a = - R^a{}_{bcd} V^b \qquad (21)$$

for any vector V^b.

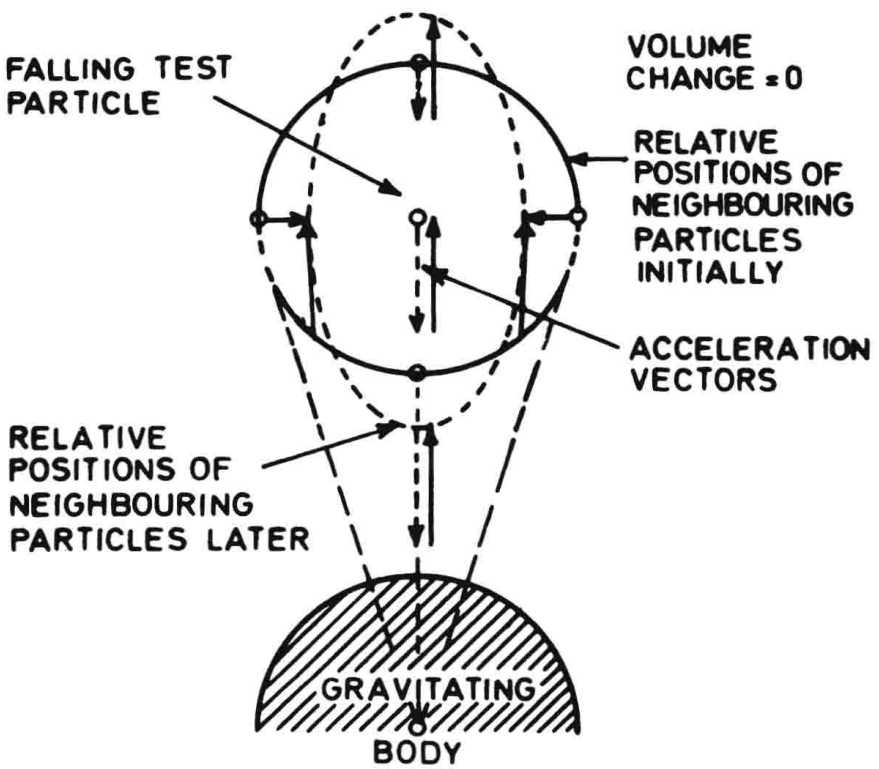


Figure 6. Distortion of sphere to ellipsoid due to tidal forces.

We shall see how this comes about later. But for the moment let us return to the Newtonian picture of tidal forces given in Figure 6. A test particle falls freely in a gravitational field, the particle being initially surrounded by a sphere of other test particles which start off at rest with respect to it. Owing to slight differences in the directions and strengths of the gravitational field, the surrounding particles will accelerate slightly with respect to the central particle, thus distorting the spherical shape into an ellipsoid. If the particles enclosed just empty space as they fall, then, as follows directly from Newton's inverse square law, the volume of the ellipsoid into which the sphere is momentarily distorted is equal to that of the sphere. But there is a decrease in the volume momentarily enclcsed as the particles enclosed as the particles accelerates inwards. This momentary decrease in volume is proportional to the total mass enclosed. In Einstien's theory the situation is almost exactly the same, a small difference arising in that the momentary decrease is now proportional not just to the mass enclosed but to the total mass plus pressures ($\propto c^{-2}$) (see Figure 7).

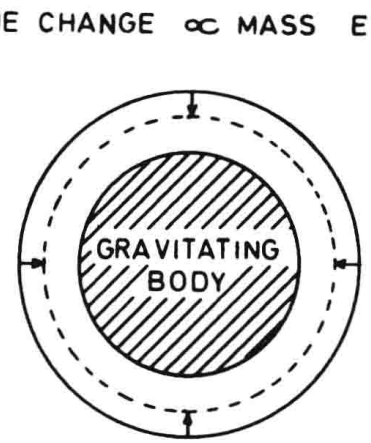


Figure 7. Volume change due to tidal forces.

To examine the general relativity situation, consider two neighbouring geodesics with tangent vectors, t^a, corresponding points being connected by the space-like position vector, p^a. Locally we could take $t^a = (1,0,0,0)$; then $t^a\nabla_a$ would be just the derivative with respect to proper-time along the geodesic. Now the geodesic equation becomes

$$\frac{d}{ds} t^a = 0 \qquad (22)$$

and the condition for the vector to continue to connect correspond-

ing points of the same pair of geodesics is

$$\underset{t}{\pounds}\, p^b = t^a \nabla_a p^b - p^a \nabla_a t^b = 0 \tag{23}$$

Thus we get the geodesic devation (or acceleration)

$$A^a = \frac{d^2 p^a}{ds^2} = (t^f \nabla_f)(t^e \nabla_e) p^a = R^a{}_{bcd}\, t^b\, p^c\, t^d \tag{24}$$

which is the Jacobi equation. Thus we see that relative acceleration is merely an effect of curvature. If we consider the trace of

$$R^a{}_{bcd}\, t^b\, t^d$$

assuming that there is no relative motion initially i.e.

$$t^a \nabla_a p^b = \frac{dp^b}{ds} = 0 \tag{25}$$

we get the momentary proportional volume change $\Delta V/V$. We define the Ricci tensor to be the first contraction of the curvature tensor

$$R_{bd} = R^a{}_{bad} \tag{26}$$

and the Ricci scalar to be the trace of the Ricci tensor

$$R = g^{bd} R_{bd}. \tag{27}$$

Thus we see that if there is no initial motion:

$$\Delta V/V = - R_{bd}\, t^b\, t^d = - R_{00} \tag{28}$$

in the geodesic's rest frame. Comparison with the Newtonian situation suggests that $-R_{00}$ should therefore be some measure of density of active gravitational mass.

Considerations of this type suggested to Einstein that the Ricci curvature of space-time must be locally definable by the distribution of matter-energy. Now, the stress-energy tensor must be conserved, i.e. divergence-free. Thus a divergence-free expression built from the Ricci tensor should be put proportional to the stress-energy tensor T_{ab}. This led Einstein to postulate the field equations

$$R_{ab} - \frac{1}{2} R\, g_{ab} = - 8\pi\, T_{ab} \tag{29}$$

The constant factor 8π comes from the requirement that the Newtonian

theory of gravitation must be approximately valid for usual planetary motion (being exactly true if we assume an infinite speed of light). In suitable co-ordinates (the eigentetrad) the stress-energy tensor is

$$T_{ab} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & p_1 & 0 & 0 \\ 0 & 0 & p_2 & 0 \\ 0 & 0 & 0 & p_3 \end{pmatrix} \tag{30}$$

Now Eq.(29) can be written as

$$R_{ab} = -8\pi \left(T_{ab} - \frac{1}{2} T\, g_{ab}\right) \tag{31}$$

whence, using Eq.(30) we have

$$R_{00} = -4\pi \left(\rho + p_1 + p_2 + p_3\right) \tag{32}$$

which gives the momentary proportional change of volume. (The effect of the pressures is small as we would have to divide them by c^2 to convert to the usual units.)

7. Gravitational Radiation and Gravitational Focusing

The gravitational field equations in vacuum i.e.

$$R_{ab} - \frac{1}{2} R\, g_{ab} = 0 \tag{33}$$

which implies that

$$R_{ab} = 0 \tag{34}$$

yield solutions corresponding to gravitational radiation. Though there is no accepted evidence that such radiation has been observed experimentally as yet, much theoretical work has been done on it. A particularly beautiful and graphic result is due to Sachs (known as the Sachs peeling theorem) which shows how the curvature tensor's geometrical structure changes along the direction of the gravitational radiation (see Figure 8). This structure is described in terms of four directions along the null cone called the gravitational principal null directions. These principal null directions (PND) can be defined as those directions in which there is no distortion of light rays by the curvature. There are precisely four of them at each general point of a general space-time. In all other directions distortion will take place. (This is a limiting case of the tidal distortion considered earlier.) In special circumstances coincidences may occur between some (or all) of the four principal

 Aspects of general relativity. *Physics and contemporary needs*, Vol. 1 (1977), 383–426.
Ed. Riazuddin

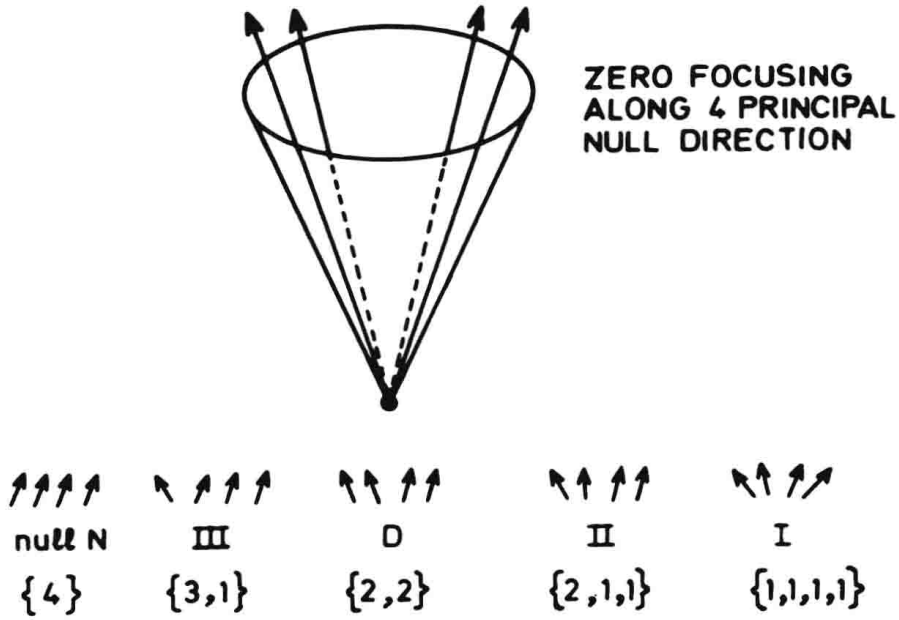


Figure 8. Sachs peeling theorem.

directions. The Sachs peeling theorem states that the part of the curvature that falls off as r^{-1} (where r is a suitable distance parameter) has all four PND pointing radially outwards, the part falling off as r^{-2} has three PND pointing radially outwards, the part falling off as r^{-3} has two PND pointing radially outwards, and so on (see Figure 8).

Bondi and Sachs have proved the theorem (called the mass-loss theorem) that outgoing gravitational radiation always decreases the mass of the system. However, it has not been adequately proved so far, that the total mass of the system must necessarily always remain non-negative, but it is conjectured that this is the case (the energy conjecture). The total mass of the system shows up in the r^{-3} term in the curvature. The total rate of mass-loss due to gravitational radiation shows up in the r^{-1} ('radiation field') terms.

The gravitational (curvature) field can be classified in terms of its PND. The classification is equivalent to what is known as the Petrov classification (see Figure 9). There are six cases to

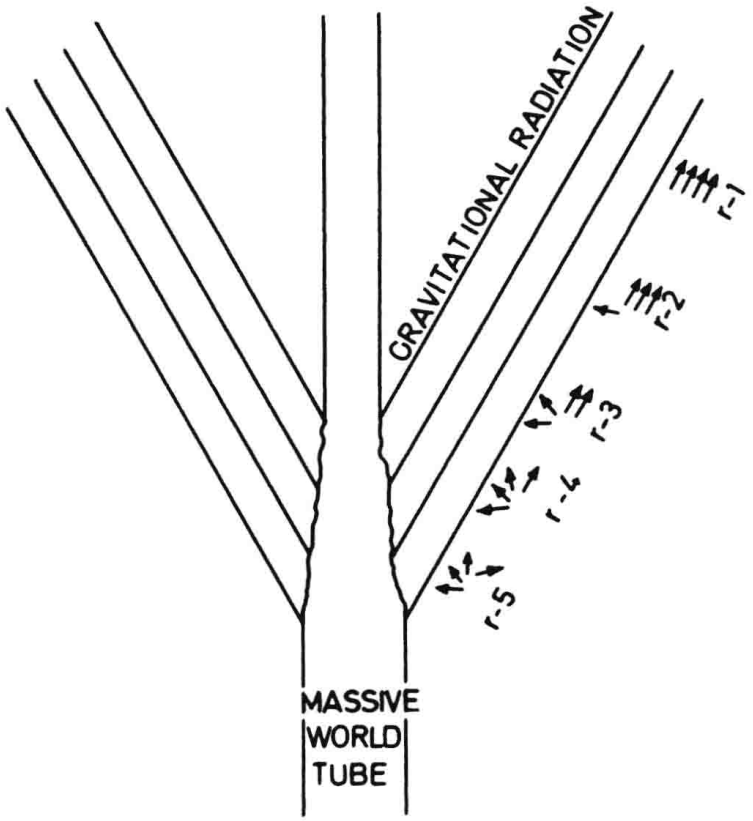


Figure 9. Principal null directions and the Petrov classification.

consider. First, there is the general case {1111} (called Petrov type I) in which the PND are distinct; then case {211} (Petrov type II) in which exactly one pair of PND coincide; then case {31} (Petrov type III) in which there is just a triple co-incidence of PND; case {22} (called type D) in which the PND coincide in pairs; case {4} (called null or Petrov type N) in which all four coincide, and finally the case {-} (or Petrov type 0) for which the curvature tensor vanishes.

The relation between PND and distortion of curvature can be obtained from the geodesic deviation equation considered earlier, the geodesics now being null (since the are light rays) i.e. the tengent vectors t^a are null

$$g_{ab}\, t^a t^b = 0. \tag{35}$$

The geodesic deviation Eq.(24) tells us that the distortion and focussing of the light rays, measured in directions orthogonal to t^a is described by

$$R_{abcd}\, q^a\, t^b\, p^c\, t^d \tag{36}$$

where p^a and q^a are chosen orthogonal to t^a (i.e. $g_{ab}\, t^a\, q^b = 0$) and $R_{abcd} = g_{ay} R^y{}_{bcd}$. The condition, in vacuum, for the null vector t^a to be directed along a PND is simply that the expression (36) should vanish for all such vectors p^a, q^a. When matter is present there is a positive net focussing effect in addition to pure distortion. This is described (compare with Eq.(32)) by the quantity

$$- R_{ab}\, t^a\, t^b \quad = \quad 8\pi\, T_{ab}\, t^a\, t^b \quad \geq \quad 0. \qquad (37)$$

The pure distortion effect would then be described by the expression (36) but where the curvature tensor is replaced by the Weyl conformal tensor (cf. Eq.(63)) and the PND are then given by these t^a's for which the resulting expression vanishes for all p^a and q^a orthogonal to t^a.

In vacuum, the net local focussing effect described by Eq.(37) is absent. There is, nevertheless, a larger-scale non-linear focussing effect which is present even in vacuum. We do not enter into the full details here, but the essential feature of this effect is one already familiar in lens optics. Imagine a single thin convex lens of power P_1. (The power of a lens is the reciprocal of its focal length.) A parallel beam of light passing through the lens will be focussed to a point at the far side of the lens. Suppose that a second such lens, of power P_2, is placed flat against the first. The combined focussing effect of the two lenses together will be that of a thin lens of power $P_1 + P_2$. But, if the two lenses are moved apart to a distance d, the power of the resulting (thick) lens will be reduced

$$P \quad = \quad P_1 + P_2 - P_1\, P_2\, d. \qquad (38)$$

The focussing effect on light rays passing through a small concentrated patch of curvature is similar to that of a lens, but in general it is a lens with <u>astigmatism</u> - that is to say, the power of the lens in one direction in its plane may differ from the power in another direction. In vacuum, the curvature is always such that the focussing is <u>purely</u> astigmatic, so that the power in any one direction in the lens is always the negative of the power in the perpendicular direction (net focussing power zero). If two such thin lenses are placed directly against one another, but rotated through a right angle with respect to each other, then their effects would exactly cancel one another out. (In each direction in the lens; the resultant power would be P+(-P) = 0.) But if the lenses are then moved apart to separation d, the resulting power would be effectively that of a positive (convex) lens of power $P^2 d$ (by Eq.(38)). Thus collections of purely astigmatic lenses can combine together in their effect to give a positive total focussing power; and in the same way, the resultant accumulative effect of different regions of vacuum curvature will again have positive focussing power. Although

in vacuum the expression (37) is zero, the curvature will exert a positive focussing effect. When matter is present and expression (37) is positive, the focussing effect is that much the greater. Similar statements apply also in the case of time-like geodesics.

One consequence of this positive focussing effect is the presence of pairs of conjugate points on null or time-like geodesics. These are pairs of points on a geodesic, at which (roughly speaking) a neighbouring geodesic to it would intersect it. A familiar example of conjugate points would be antipodal points on a great circle on a sphere. An arc of a great circle which contains no pair of conjugate points is the shortest arc connecting its end points. But if an open great circle arc extends so far about the sphere that it contains a pair of antipodal points, then it will not be the shortest arc joining the end points. This illustrates an important general property of conjugate points in Riemannian geometry: minimal arcs are geodesics free of pairs of conjugate points (except possibly at their end points). In the case of time-like geodesics in space-time, we are concerned not with shortest arcs but with world lines of maximum proper-time. If p and q are two space-time events connected by a time-like arc α of maximum proper-times then the arc α is a time-like geodesic containing no pair of conjugate points (except possibly its end points).

An illustration of this phenomenon can be given in terms of motions in free orbit about the sun. The sun's gravitational field is assumed to be spherically symmetric and, in the region external to the sun itself, to be given by a solution of Einsteins' vacuum field equations. This entails that, in suitable co-ordinates (θ and ϕ being normal spherical polars, t being a time co-ordinate and r a radial co-ordinate scaled such that the surfaces t = constant and r = constant are intrinsically spheres of total surface area $4\pi r^2$), the space-time metric takes the Schwarzschild form

$$ds^2 = \left(1 - \frac{2m}{r}\right) dt^2 - \frac{dr^2}{\left(1 - \frac{2m}{r}\right)} - r^2 d\theta^2 - r^2 \sin^2 d\phi^2. \qquad (39)$$

Now imagine two small bodies P and Q, where P is held stationary in the given co-ordinates (i.e. having a world line for which r, θ and ϕ are constant, say r_0, $\pi/2$, 0) and where Q moves freely in a circular orbit about the sun, starting in coincidence with P at an event a and returning to coincidence with Q at an event b (i.e. having a geodesic world-line with, say, $r = r_0$, $\theta = \pi/2$, and ϕ varying from 0 to 2π), (see Figure 10). It is evident from Eq.(39) that Q experiences a shorter proper-time interval between a and b than does P (since the term in $-r^2 \sin^2\theta \, d\phi^2$ contributes negatively) even though Q has a geodesic orbit and not P. (The body P would have to

 Aspects of general relativity. *Physics and contemporary needs*, Vol. 1 (1977), 383–426. Ed. Riazuddin

fire rockets continually in the direction of the sun in order to maintain its position.) Recall that locally a time-like geodesic must maximize the time interval between any two of its points. The reason that Q's world line does not maximize the interval between a and b is that it contains pairs of conjugate points (i.e. given by $\phi = \phi_0$, $\phi = \phi_0 + \pi$, where $0 \leq \phi_0 < \pi$). The actual maximum proper-time interval between a and b would be achieved by a third body R, which moves directly out from the sun from a at just such a speed that it falls freely back to meet P and Q just as they come into coincidence at b (see Figure 10). The world-line of R would be a time-like geodesic free of pairs of conjugate points between a and b.

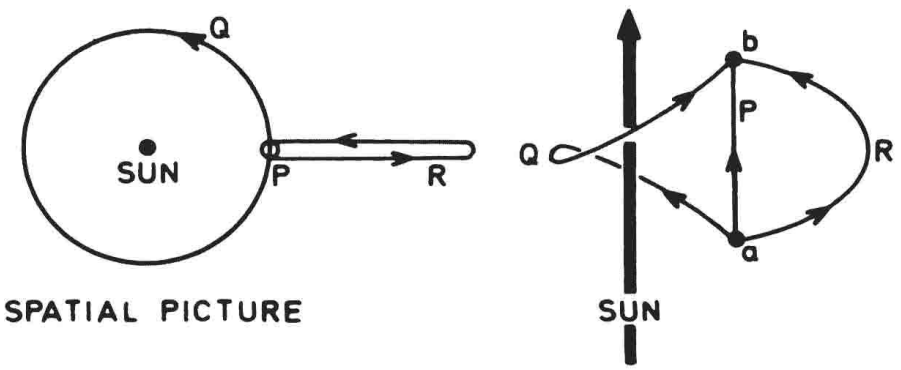


Figure 10. Space-time picture.

8. Co-ordinate and Essential Singularities

We notice that the expression for the metric given in Eq.(39) is singular for r = 0 and for r = 2m. When r > 2m the metric gives a perfectly good description of a static vacuum field, but when r ≤ 2m the expression for the metric (Eq.(39)) does not make physical sense because it is singular across r = 2m. The question can be asked whether the singularity is physically significant, however, and the same question arises at r = 0. We know that in ordinary two-dimensional Euclidean space, the origin would appear to be singular if polar co-ordinates are used, whereas in fact the origin is geometrically no different from any other point. Such an apparent singularity is called a co-ordinate singularity, characterized by the fact that it can be removed by a change in co-ordinate systems (e.g. in this case by returning to standard Cartesian co-ordinates). The question we are asking is "Are the singularities at r = 0 and r = 2m just co-ordinate singularities or are they due to the essentially singular nature of the physical situation being described?" If we transform the time parameter, t, to either the retarded time of Eddington and Finklestein

$$u = t - r - 2m \log(r - 2m) \qquad (40)$$

or their advanced time

$$v = t + r + 2m \log(r - 2m) \qquad (41)$$

we can express the metric as

$$ds^2 = \left(1 - \frac{2m}{r}\right) du^2 + 2dr\, du - r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \qquad (42)$$

or

$$ds^2 = \left(1 - \frac{2m}{r}\right) dv^2 - 2dr\, dv - r^2(d\theta^2 + \sin^2\theta\, d\phi^2). \qquad (43)$$

The picture in these co-ordinates looks as shown in Figures 12a and 12b respectively instead of as in Figure 11 (which applies for the usual spherical polar co-ordinates).

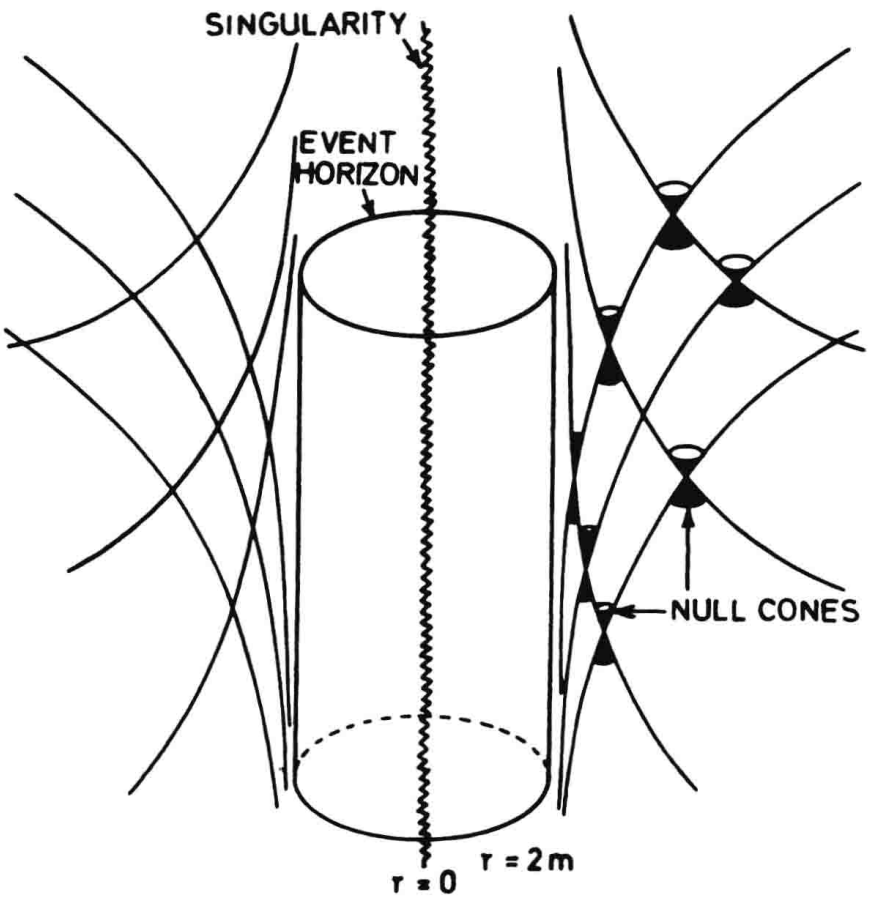


Figure 11. Radial null geodesics in Schwarzschild co-ordinates

 Aspects of general relativity. *Physics and contemporary needs*, Vol. 1 (1977), 383–426.
Ed. Riazuddin

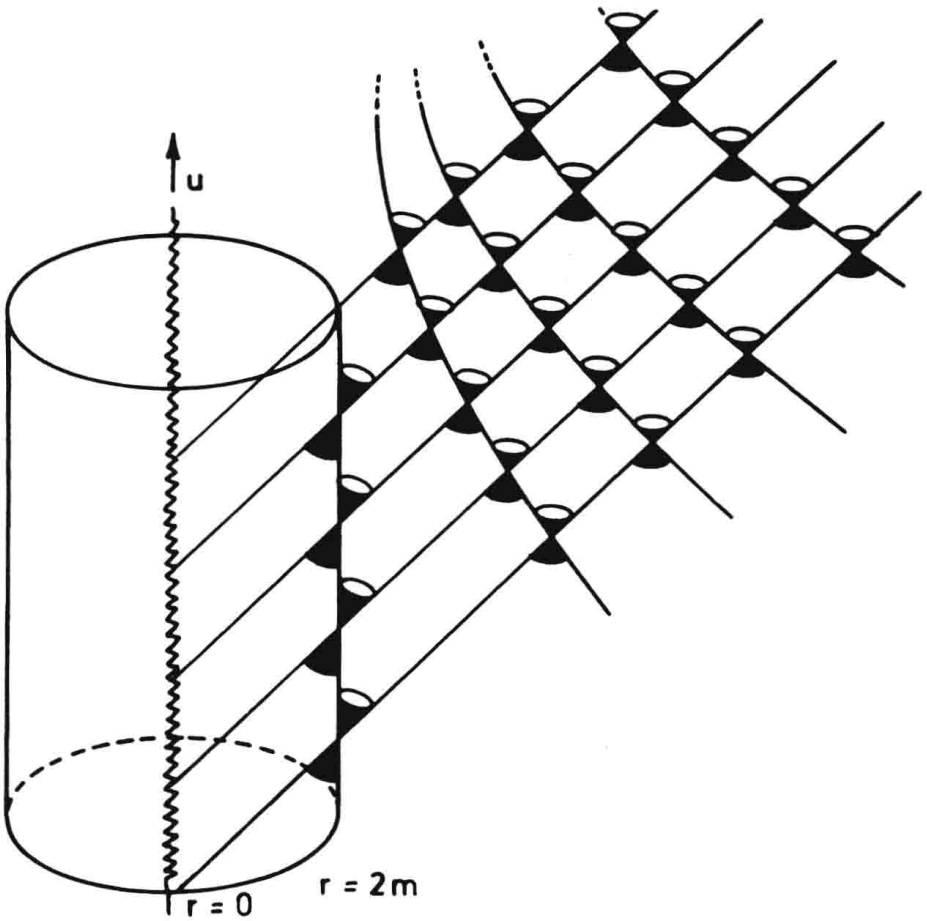


Figure 12a. Radial null geodesics in retarded co-ordinates

We find that there is no singularity in Eqs.(42) and (43) at $r = 2m$, showing that $r = 2m$ is only a co-ordinate singularity of the original expression (Eq.(39)) for the metric. However, it is not possible to find co-ordinates in which the singularity at $r = 0$ is removed. This is because the space-time curvature (as measured in some appropriately invariant way, e.g. in terms of the scalar $R^a{}_{bcd}\, R^b{}_{aef}\, g^{ce}\, g^{df}$) is unbounded as $r = 0$ is approached.

It had been noticed, at the end of the eighteenth century, by Laplace, that any body of diameter less than $r = 2m$ (in gravitational units) must be invisible from large distances, even in an appropriately interpretted version of Newtonian theory, because the escape velocity at the surface of the body would exceed the velocity of light. This assumes that the velocity of light at the surface of the body is the same as at infinity. Such an assumption may seem

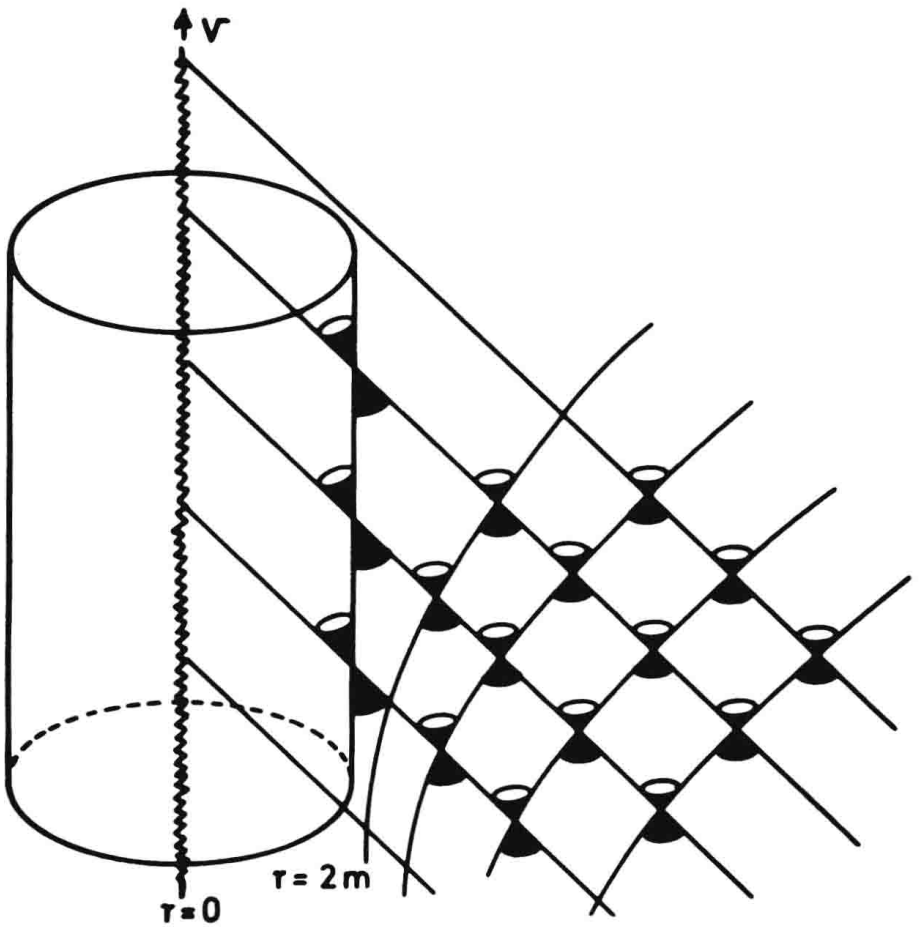


Figure 12b. Radial null geodesics in advanced co-ordinates

unnatural in Newtonian theory but is natural enough in a relativistic theory. Thus, it is not surprising that the metric given in Eq.(43) should have the property that light can not escape from within the r = 2m. The space-time locus described by r = 2m is called the (absolute) event horizon - this being a surface from within which no information can come out (i.e. get to infinity).

It has been conjectured that any singularity, arising naturally in general relativity, must have an event horizon around it. This is called the cosmic censorship conjecture. It is possible, however, that this is not valid and naked singularities, visible from infinity, can occur (hence the term 'censorship'). Alternately it could be that the event horizon is there to start with but disappears later - in which case both an event horizon and a naked singularity would be present.

 Aspects of general relativity. *Physics and contemporary needs*, Vol. 1 (1977), 383–426. Ed. Riazuddin

It has been proved that a space-time manifold is necessarily singular (cf Hawking and Ellis) if (1) it satisfies Einstein's equation with the strong energy condition (i.e. $R_{ab}\, t^a\, t^b \leq 0$ for all time-like vectors t^a), (2) it contains no closed time-like curves (causality), (3) it is suitably generic and (4) it contains either (a) a closed space-like hypersurface, or (b) a trapped surface (i.e. a closed space-like 2-surface with the property that the null geodesics orthogonal to it have positive net convergence into the future at every point of the surface) or (c) a point whose light cone begins to converge somewhere on every generator of the cone, as shown in Figure 13. The singularity may come about due to matter collapsing under gravity (Figure 14) for example.

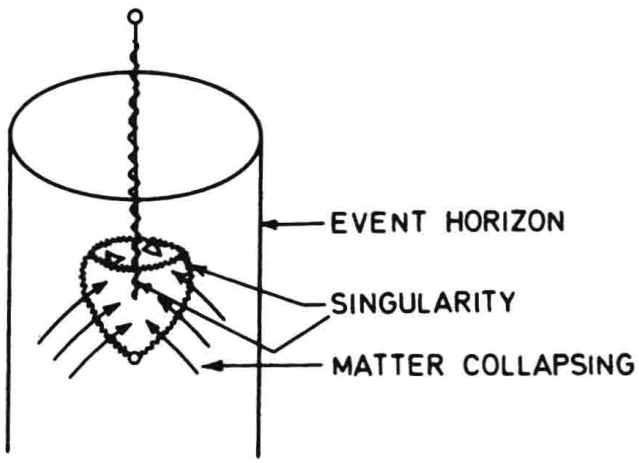


Figure 13. Converging light cone.

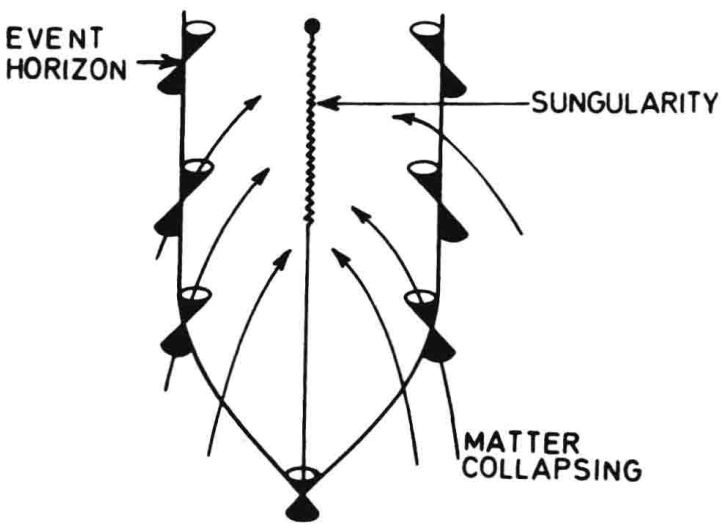


Figure 14. Collapse under gravity.

9. Black Holes

When a body collapses within its event horizon the resulting configuration is called a black-hole, as no light can get out of it. For an electrically neutral non-rotating body, all irregularities in the external field will eventually be dissipated away as gravitational radiation when the body collapses and the configuration will end up as a Schwarzschild black hole, the vacuum metric being given by Eq.(43). From outside the event horizon it will behave like an energy sink, only showing its effect through its gravitational field and the black-hole's propensity to absorb energy. The extent of the black-hole would be a sphere centred at the singularity, with radius 2m (m being the mass of the black hole). If a test particle were sent into the black-hole it would appear, to an observer at infinity, as hovering at the edge of the black-hole, rapidly growing redder and fainter until it faded completely away, i.e. the light from it would be increasingly red-shifted. If we were to follow the test particle in and study light signals sent by it, they would appear as shown in Figure 15a, or, in space-time terms as shown in Figure 15b.

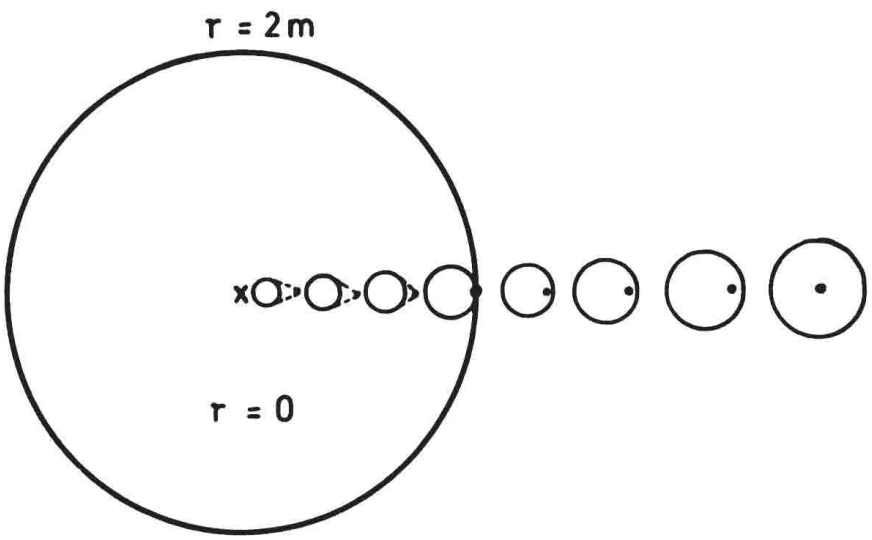


Figure 15a. Light signals in a non-rotating black-hole situation.

We may also envisage a charged black-hole. The metric for this type of black-hole for a spherically symmetric body of mass m and charge q (known as the Reisener-Nordstrom metric) is

$$ds^2 = \left(1 - \frac{2m}{r} + \frac{q^2}{r^2}\right) dt^2 - \frac{dr^2}{\left(1 - \frac{2m}{r} + \frac{q^2}{r^2}\right)} - r^2(d\theta^2 + \sin^2\theta d\phi^2) \qquad (44)$$

 Aspects of general relativity. *Physics and contemporary needs*, Vol. 1 (1977), 383–426.
Ed. Riazuddin

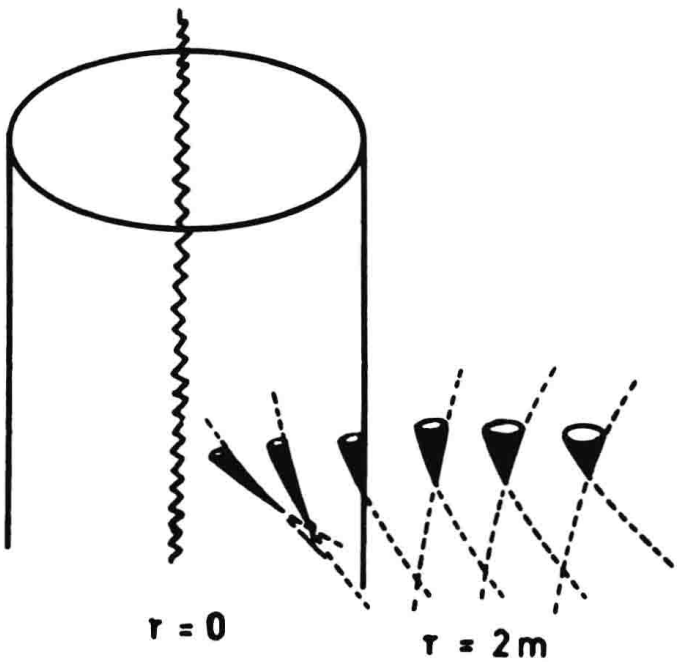


Figure 15b. Light cones in a non-rotating black-hole.

or in advanced co-ordinates

$$ds^2 = \left(1 - \frac{2m}{r} + \frac{q^2}{r^2}\right) dv^2 - 2dr\, dv - r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \tag{45}$$

Eq.(44) has co-ordinate singularities for

$$r = m \pm \sqrt{m^2 - q^2} = r_{\pm} \tag{46}$$

while the surfaces are regular for the form given by Eq.(45), the event horizon being $r = r_+$. The metric given in Eq.(44) clearly reduces to the Schwarzschild metric (Eq.(39)) in the case $q = 0$. In any realistic situation we have $m >> q$ and the effect of the charge on the field can be ignored.

Alternatively, we could envisage a rotating black-hole, of mass m and angular momentum a. Provided $m > a$ we can use the metric

$$ds^2 = dt^2 - (r^2 + a^2\cos^2\theta)\left(\frac{dr^2}{r^2 - 2mr + a^2} + d\theta^2\right) - (r^2 + a^2)\sin^2\theta\, d\phi^2 - \frac{2mr}{(r^2+a^2\cos^2\theta)}\left(\frac{a^2}{m^2}\sin^2\theta d\phi - dt\right)^2 \tag{47}$$

called the Kerr metric. (More generally, we can incorporate both a charge q and an angular momentum am. The resulting metric is referred to as the Kerr-Newman metric.) Israel, Carter, Hawking and D.Robinson have proved results indicating that this is the only final stable solution and that any irregularities in the initial gravitational field will ultimately be carried away by gravitational radiation.

The Kerr metric (Eq.(47)) has, not only the co-ordinate singularities at

$$r = m \pm \sqrt{m^2 - a^2} = r_{\pm} \tag{48}$$

but also a 'ringlike' singularity at r = 0. The r = 0 singularity is a curvature singularity (as in the Schwarzschild and Reissner-Nordstrom cases), with $r = r_{\pm}$ being co-ordinate singularities, removable by making the co-ordinate replacements $(t,r,\theta,\phi) \to (v,r,\theta,\chi)$ for which

$$dv = dt + \frac{(r^2 + a^2)}{r^2 - 2mr + a^2}\, dr \tag{49a}$$

$$d\chi = d\phi + \frac{a}{r^2 - 2mr + a^2}\, dr \tag{49b}$$

thereby yielding the metric

$$ds^2 = dv^2 - 2dr(dv - a \sin^2\theta\, d\chi) - \frac{2mr}{(r^2 + a^2 \cos^2\theta)}$$

$$(dv - a \sin^2\theta\, d\chi)^2 - (r^2 + a^2 \cos^2\theta)\, d\theta^2. \tag{50}$$

There is a region (called the ergosphere) namely

$$r_+ < r < m + (m^2 - a^2 \cos^2\theta)^{\frac{1}{2}} \tag{51}$$

in which no physical particle can remain stationary as viewed from infinity, since, to do so, it would exceed the local light velocity. This is because, in a sense, the rotational field of the Kerr black-hole causes the space to be 'whirled round' with it - the effect of dragging inertial frames. The singularities and the ergosphere are pictured in space-time in Figure 16a, and in space in Figures 16b and 16c.

Notice that if a > m then the event horizon disappears (from Eq.(48)). It is not clear what the significance of the Kerr metric is, in this case. In fact, it describes a naked singularity, but there is no reason to attach any special significance to this. Whether or not a rapidly rotating body would every collapse to such a

 Aspects of general relativity. *Physics and contemporary needs*, Vol. 1 (1977), 383–426. Ed. Riazuddin

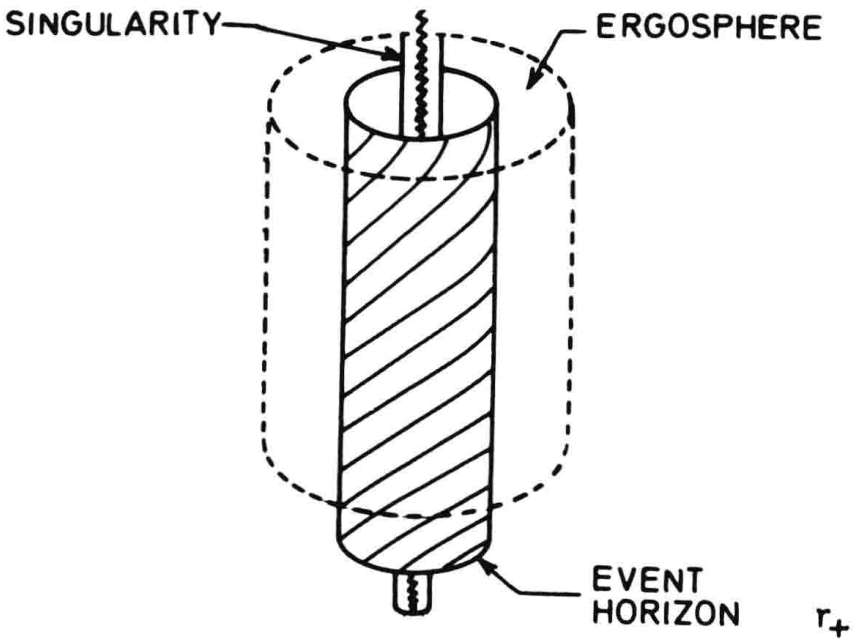


Figure 16a. Space-time picture of the Kerr black-hole.

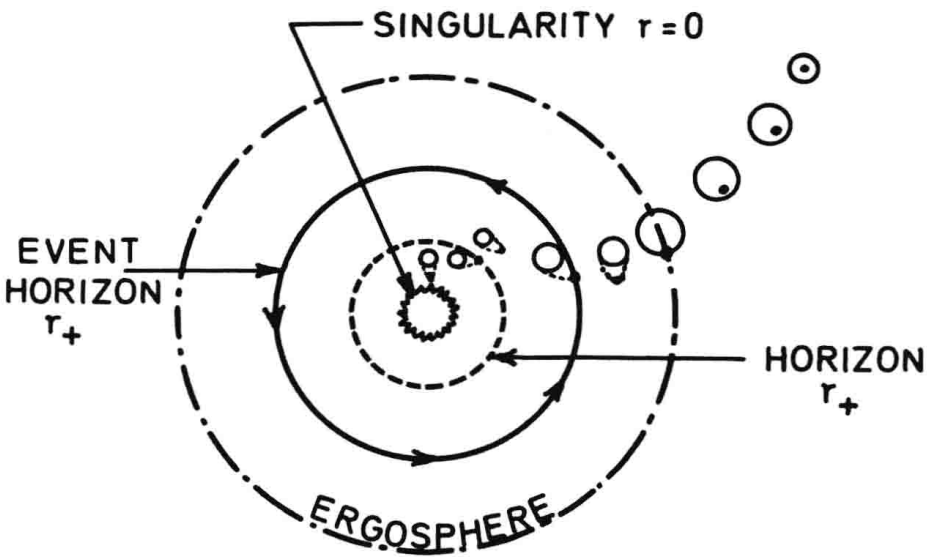


Figure 16b. Space picture of Kerr black-hole (viewed along rotation axis).

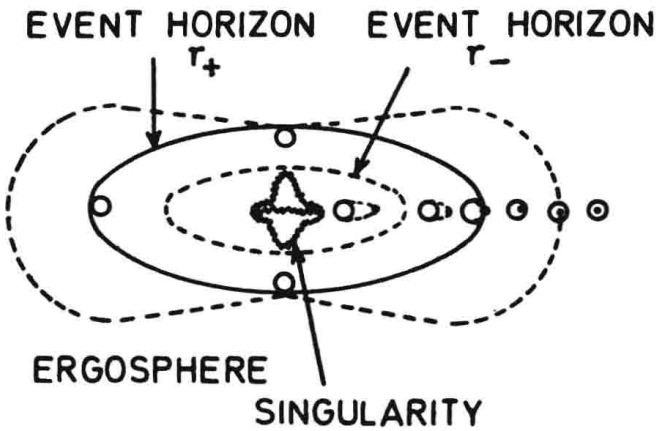


Figure 16c. Edge-on view of Kerr black-hole.

configuration is not known. If Cosmic Censorship is true, then it would certainly not do so.

The surface area of the event horizon of a Schwarzschild black-hole is

$$A = 4\pi(2m)^2 = 16\pi\, m^2. \qquad (52)$$

For a Kerr black-hole it is

$$A = 8\pi\, m(m + \sqrt{m^2 - a^2}\,) = 8\pi\, mr_+ \qquad (53)$$

There is a general principle (known as the area principle)which states that the surface area of the absolute event horizon of a black-hole is non-decreasing with time. The first published precise theorem to this effect was given by Hawking (cf Hawking and Ellis). The Cosmic Censorship Hypothesis needs to be assumed for this.

Eq.(52), together with the area principle, shows that if a stationary non-rotating (i.e Schwarzschild) black-hole absorbs some material and then settles down to become stationary and non-rotating again then the final mass of the hole must be greater than its initial mass. Thus, it is not possible to extract energy (=mass) from a non-rotating black-hole by lowering objects into it and extracting their gravitational potential energy in some way. However, if angular momentum is present, the area principle is not so restrictive. In Eq.(53) it is possible to increase A and decrease m at the same time, provided a is also decreased sufficiently. This suggests that it might be possible to extract rotational energy from a rotating black-hole. In fact this turns out to be the case: a small body, b_1, is thrown into the ergosphere, where it splits into two

 Aspects of general relativity. *Physics and contemporary needs*, Vol. 1 (1977), 383–426. Ed. Riazuddin

bodies, b_2 and b_3, such that b_2 falls into the black-hole carrying both negative angular momentum and negative energy, the body b_3 being ejected with more total energy than that of the original body b_1. For a test particle of four momentum p^a, we can define the total energy E by

$$E = p^a k_a \tag{54}$$

where k_a is the Killing vector describing the stationarity of the space-time, a Killing vector being one which satisfies the Killing equation

$$\nabla_a k_b + \nabla_b k_a = 0. \tag{55}$$

Along a geodesic ($p^a \nabla_a p^b = 0$) E is conserved. The angular momentum is similarly defined but where k^a now describes a rotational space-time symmetry. The difference between the total energy (kinetic plus potential plus rest energy) of b_3 and that of b_1 represents the energy extracted from the black-hole (see Figure 17). This comes from the rotational energy of the hole, since its angular momentum is likewise reduced.

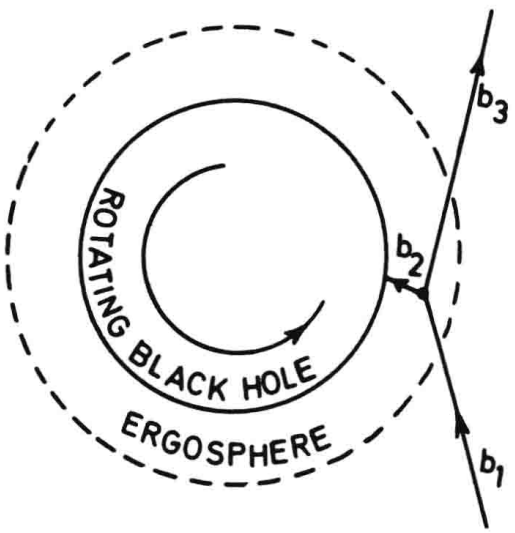


Figure 17. Energy extraction from a rotating black-hole.

10. The Area Principle

As we have just seen, the area principle allows the area of a black-hole to increase but not to decrease - and accordingly there is an irreducible mass

$$m_I = \sqrt{A/16\pi} \tag{56}$$

for any black-hole of area A, namely that mass for which a Schwarzschild black-hole would have area A (as given by Eq.(52)). Any rotating black-hole of area A must have a greater mass than m_I (by Eq.(53)). Hawking showed that the area principle can also be applied to colliding black-holes: the total area of two intersecting black-holes is less than the area of the final black-hole formed when the two incoming ones merge. (The fact that the result of such a collision is a black-hole and not a naked singularity is a consequence of the assumption of cosmic censorship.) This version of the area principle is expressed diagramatically in Figure 18a. The merging of the black-holes can be seen diagramatically in terms of the space-like sections of the black-holes at different times, as shown in Figure 18b.

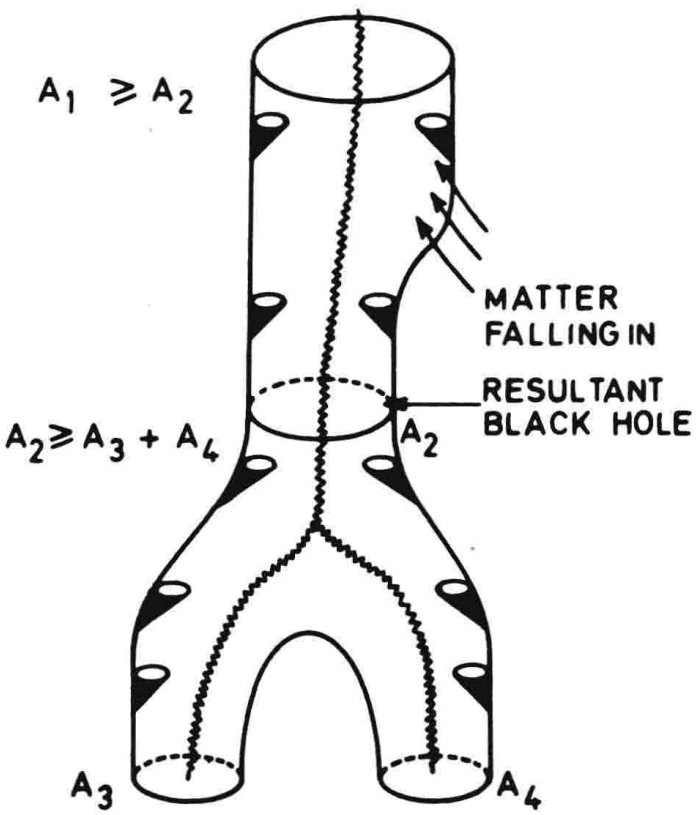


Figure 18a. The area principle.

To see how the proof of the area principle works, we invoke the following formula for the propagation of an element of cross-sectional surface area δA of the event horizon (which is a null hypersurface) along its null geodesic generators

$$\frac{d^2(\delta A)^{\frac{1}{2}}}{du^2} = -(\delta A)^{\frac{1}{2}} (|\sigma|^2 + \Phi) \qquad (57)$$

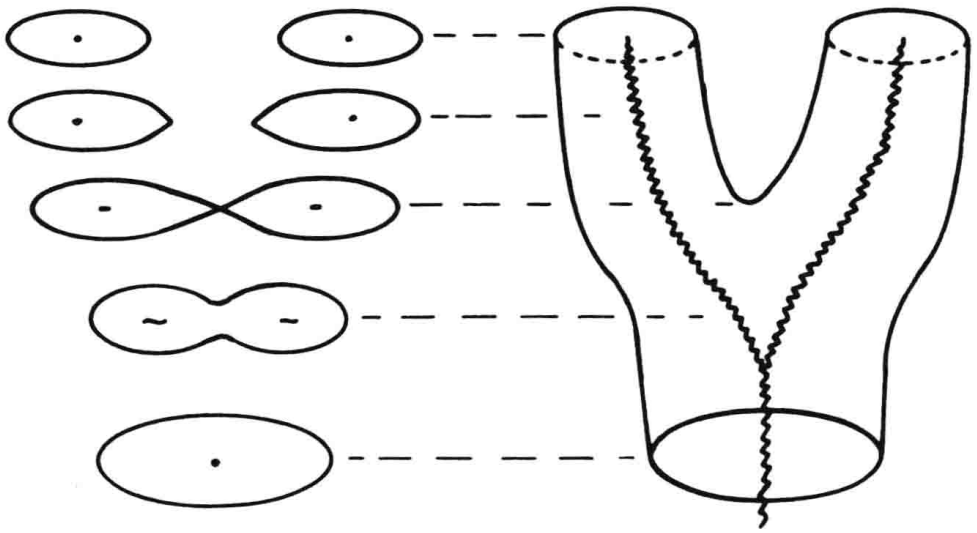


Figure 18b. Merging of black-holes.

Here u is an affine parameter (cf Eq.(18)) on the null geodesic, σ is a complex quantity measuring the shearing of the neighbouring null geodesics and is the Ricci tensor component

$$\Phi = -\frac{1}{2} R_{ab} t^a t^b = 4\pi T_{ab} t^a t^b \geq 0 \qquad (58)$$

t^a being tangent to the null geodesic. From Eqs.(57) and (58) it follows that

$$\frac{d^2(\delta A)^{\frac{1}{2}}}{du^2} \leq 0. \qquad (59)$$

Thus if δA were to decrease at any stage it must become zero after a finite affine parameter 'time' has elapsed (see Figure 19). This is impossible if the Cosmic Censorship Conjecture is true. It follows that δA can never decrease along any null generator of the absolute even horizon. In addition, it is possible for new generators to emerge along the absolute event horizon, but not to disappear. Thus the area of cross-section of the horizon can increase but never decrease, provided cosmic censorship is true and the classical inequality (58) holds (weak energy condition), (see Figure 20).

When quantum effects are involved, the inequality (58) need no longer be true, and the area principle can be violated. This is the situation, in particular, with the Hawking black-hole evaporation process, according to which a black-hole emits thermal radiation, by quantum mechanical effects, at a temperature inversely, propor-

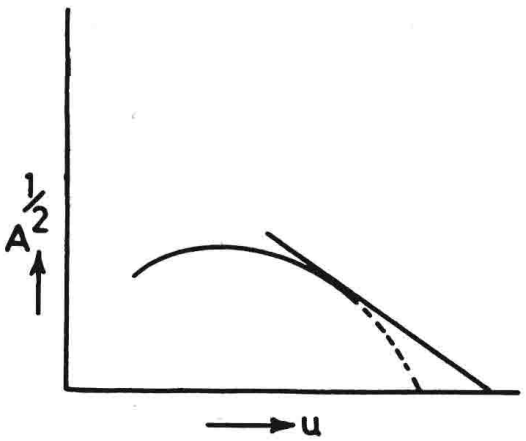


Figure 19. Rough proof of Area Principle.

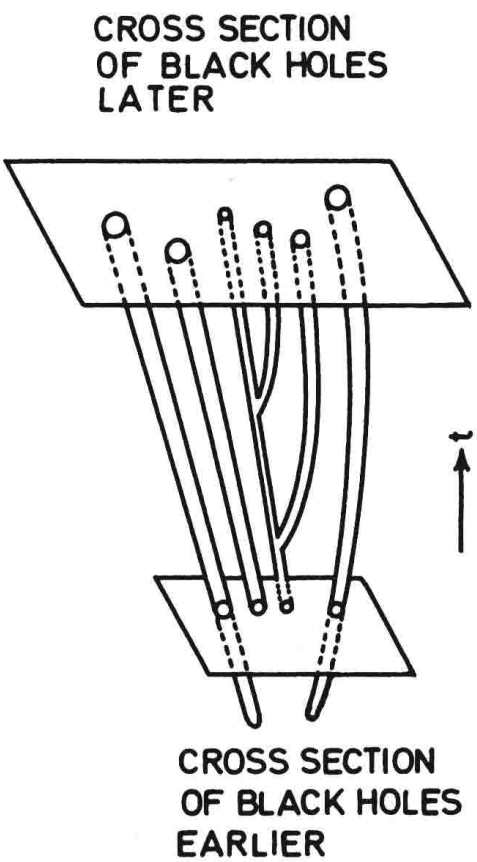


Figure 20. Rough geometric proof of Area Principle.

tional (in the non-rotating black-hole case) to the black-hole's mass. A rough intuitive derivation of this phenomenon has been suggested by Hawking. In the neighbourhood of the event horizon a virtual pair of particles is produced, one of these having positive energy and the other negative energy. The negative energy particle tunnels to within the black-hole where it can exist as a real particle to escape to infinity. The rate at which this process takes place is governed by the degree of tipping of the light cones (cf Figure 12a) at the event horizon. This degree of tipping is described by a quantity κ called the surface gravity of the black-hole which, for the Schwarzschild and Kerr metrics respectively, takes the values

$$\kappa = \frac{1}{4m} \quad \text{and} \quad \kappa = \frac{4\pi}{A}\sqrt{m^2 - a^2} \tag{60}$$

(cf Eq.(53)). Hawking has shown that the temperature T of black-hole is given by

$$T = \frac{h\kappa}{2\pi} \tag{61}$$

(choosing temperature units so that Boltzman's constant is equal to one). Correspondingly, the entropy of the hole is

$$S = \frac{1}{4h} A \tag{62}$$

a formula which (apart from the precise assignment of the numerical value ¼) had been previously suggested by Bekenstein.

Bekenstein's argument for the assignment of an entropy in proportion to the black-hole's surface area is basically a thought experiment. One attempts to violate the second law of thermodynamics by lowering an object of high entropy into the black-hole, while attempting to extract its energy content in a low entropy form at infinity. In fact, if the back reaction of the black-hole geometry is ignored, an effective violation of the second law can be achieved in this way. But by the area principle, if the geometry of the black-hole is disturbed in any way, its surface area must increase. This suggests a functional relation between the black-hole's entropy and its surface area, which can be expected to be linear because of the addit ve properties of entropy, as applied to the situation of a pair of widely separated black-holes, the appearance of Planck's constant coming from dimensional considerations. Bekenstein's detailed argument yields an expression of the form (62). The argument is incomplete without the Hawking radiation, but when this effect is incorporated as well, a remarkably consistent overall picture is obtained.

It should be emphasized, however, how very minute the Hawking temperature is for any black-hole that could arise in stellar collapse. For a black-hole of the mass of the sun, the temperature would

be less than 10^{-7} °K. Only for very minute black-holes could the temperature be significant. And it would seem that such black-holes could only arise if they were produced in the big bang - and then only if this big bang were of a sufficiently chaotic nature (a possibility which may be thought of as unlikely, in view of the arguments of the next section). But if such mini black-holes do occur, the Hawking radiation could be a substantial effect. This radiation would carry away mass from the hole and cause its surface area to be reduced - in violation of the area principle! This shows that, in effect, the area principle cannot be expected to hold when quantum effects are being considered. These quantum effects can provide an effective violation of the weak energy condidition (cf. Eq.(58)) on which the area principle was based. Then as the black-hole is reduced in mass its temperature goes up (cf Eqs.(60) and (61)), its radiation correspondingly increases and a runaway situation ensues, with the black-hole (presumably) disappearing in one final explosion with a naked singularity being momentarilly produced at the end. (So cosmic censorship also appears to be violated in such a quantum mechanical regime.)

11. Speculations on Cosmology

It is normally taken to be a basic law of local physics that time-reversal invariance holds - at least as regards those physical laws that are significant in governing the vast majority of observed phenomena (the one exception arising in K^0-decay). Thus, if any elementary process such as

$$A + B \longrightarrow C \longrightarrow D + E + F$$

is allowed, then so also is

$$D + E + F \longrightarrow C \longrightarrow A + B$$

However, there is a large scale statistical time-asymmetry in physical processes which is embodied in the principle of increase of entropy with time. Thus a cup falling will break into little pieces, but the little pieces will not come together to form the cup on their own. This statistical asymmetry is attributed not to an asymmetry in the local physical laws, but to a difference in nature between the boundary conditions in the remote past and remote future.

Now, if we consider a K = +1 Friedmann-type cosmology, we find that the past and future boundary conditions of the universe are both space-time singularities. Thus the presence of statistical time-asymmetry in the universe implies that the past singularity must differ in structure from the future singularity. Even in K = -1 Friemann-type cosmologies, a similar distinction must occur, this time between the initial big-bang singularity and the singularities

occurring inside black-holes (since these form the final boundary conditions of observers who fall into black-holes). Initial singularities must therefore be, in some sense, low entropy ones. So the question arises as to how one might characterize the geometry of a singularity.

To deal with this question consider how the Riemann curvature tensor R_{abcd} splits into the Weyl (or conformal) curvature tensor, C_{abcd} and the Ricci part, R_{ab}, of the Riemann curvature tensor

$$R_{abcd} = C_{abcd} + \frac{1}{2}(R_{ac}\, g_{bd} + R_{bd}\, g_{ac} - R_{ad}\, g_{bc} - R_{bc}\, g_{ad}) - \frac{1}{6} R(g_{ac}\, g_{bd} - g_{ad}\, g_{bc}) \qquad (63)$$

The Weyl tensor possesses the characteristic symmetry properties of the Riemann tensor, but, in addition, it satisfies the 'vacuum' condition

$$g^{ac}\, C_{abcd} = 0. \qquad (64)$$

We may regard the Weyl tensor as describing the free gravitational part of the curvature. It is responsible for the pure 'volume preserving' part of the tidal force (cf Figure 6) while the Ricci part of the Riemann tensor is the part directly defined by the matter (via Einstein's field equations) and gives rise to a 'volume reducing' tidal effect. If C_{abcd} vanishes the space is said to be conformally flat, i.e. the metric can be expressed locally as

$$ds^2 = F(t,x,y,z)(dt^2 - dx^2 - dy^2 - dz^2) \qquad (65)$$

where $F(t,x,y,z)$ is a smooth positive function.

In the case of spherically symmetric gravitational collapse to a singularity we can easily see that Weyl curvature becomes infinite but the Ricci curvature is zero in most places as the singularity is approached (being non-zero only on a small portion of the singularity, namely that which is reached by the collapsing matter). Furthermore, in asymmetric collapse, we expect the Weyl curvature to dominate over Ricci curvature in the process of generic collapse to a singularity, as shown, in particular, by Belinskii, Lifshitz and Khalathikov's analysis of 'generic' singularities, or by various other arguments.

For the big-bang, in the case of the exact Friedmann universe (which is completely isotropic), the Weyl tensor vanishes (since the principal null directions must be undefined) implying that the space-time is conformally flat. Here it is clear that the Ricci curvature is infinite and the Weyl curvature zero as the singularity

is approached. This is opposite to the situation of gravitational collapse to a singularity and it may be postulated that a similar structure for the big-bang singularity should obtain even when the exact symmetry of the Friedmann model is not assumed.

It appears, then, that in the initial low entropy situation the Weyl tensor is very small, while in the later high entropy situations the Weyl tensor becomes very large. Such an association of entropy with the Weyl tensor makes sense provided it is the entropy in the gravitational field that is being referred to here. With gravitation, it is 'clumped' states that have the high entropy and smoothed-out ones that have low entropy (owing to the attractive unstable nature of the gravitational interaction) and the more smoothed-out the state, the smaller will be the average magnitude of the Weyl tenscr. At present, this can only be understood in a qualitative way. There is no precise quantity known that can be consistently interpreted as 'the entropy in the gravitational field'. Furthermore, there is the unusual aspect of this that one is attempting to identify some quantity of geometrical measure of the space-time with a statistical quantity, namely entropy. But this aspect of the identification, though perhaps surprising at first, can not be regarded as an objection, in view of the fact that it is already present in the Bekenstein-Hawking expression (62) for the entropy of a black-hole.

Let us thus make the hypothesis that something in the local physical laws forces the Weyl curvature to be zero (or almost zero) near any initial singularity in the space-time. The fact that there is something time-asymmetrical somewhere in local physics shows up in the time-reversal symmetry-breaking of the K^{o}-decay. (But, in addition to T-violation, we would need TCP-violation. Maybe we shall sometime find a TCP-violating phenomenon in local physics.) The vanishing (or small) initial Weyl tensor would force the initial entropy in the gravitational field to be low, while allowing maximum entropy (i.e. thermal equilibrium) for all the matter fields - this being the normal supposition in the big-bang. Then, as the universe evolves, the gravitational field would gradually take up entropy, allowing stars to shine and life to evolve (see Figure 21) - the entropy reaching its maximum value in the Weyl curvature divergence in the final singularities.

If this hypothesis is true, there would be the following indications:-

(1) Black-holes can exist but 'white-holes' (i.e. time-reversed black-holes) cannot;

(2) A big-bang universe will be highly isotropic and homogeneous;

(3) There would probably not be any primordial or mini-black-

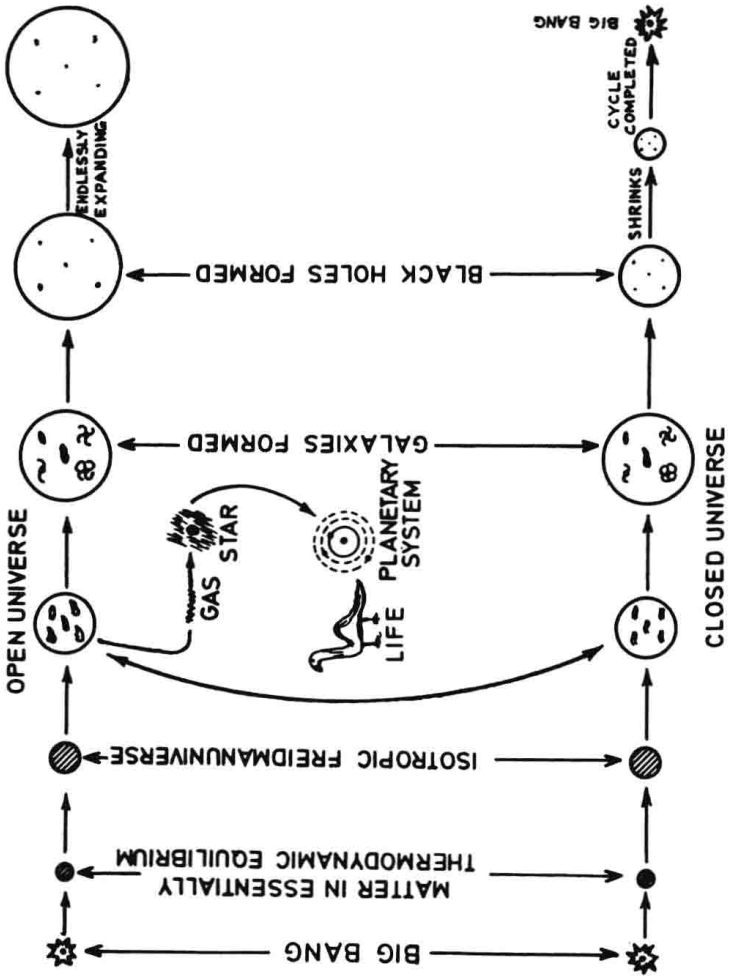


Figure 21: The gradual taking up of entropy into the gravitational field throughout the history of the Universe.

holes;

(4) One would probably expect TCP-violation in some local experiments;

(5) There would be statistical symmetry compatible with that which is observed.

Bibliography

1. G.F.R.Ellis and S.W.Hawking, 'The Large Scale Structure of Space-Time', Cambridge University Press (1973).

2. R.Penrose, 'Structure of Space-Time' in 'Battelle Rencontres', Ed. C.M.De Witt and J.A.Wheeler, Benjamin, New York (1968).

3. C.W.Misner, K.S.Thorne and J.A.Wheeler, 'Gravitation', W.H.Freeman (1973).

4. W.Rindler, 'Essential Relativity', von Nostrand, Reinhold (1969).

92
IS NATURE COMPLEX?

The Encyclopaedia of Ignorance. Eds R. Duncan and M. Weston-Smith. Pergamon, Oxford, 1977

Roger Penrose, F.R.S.

Rouse Ball Professor of Mathematics and Fellow of Wadham College, Oxford.

Awarded the Dannie Heineman Prize of the American Physical Society and the American Institute of Physics, 1971, the Eddington Medal of the Royal Astronomical Society of London, 1975 and the Adams Prize, Cambridge University 1967.

IS NATURE COMPLEX?

1. NATURE AND NUMBER

Though Nature is undoubtedly subtle, she is surely not malicious. This, at least, we have on the authority of Einstein.†[1] But is she complex? The word has more than one meaning: the first, nearer to "complicated" than "subtle"; and the second, a mathematical meaning concerning the nature of number. It is this second meaning that forms the main subject of this essay, though there will be relevance also to the first.

That Nature can be usefully described, at least to a considerable degree, according to the laws of number, has been in evidence for many centuries. But what is not so familiar to those without a mathematical background is that there are several *different kinds* of number, many of which are nevertheless subject to the same arithmetical laws.

Most primitive are the so-called *natural numbers*:

$$0, 1, 2, 3, 4, \ldots$$

and just about anything in our world can be quantified with their aid. We may speak of 3 apples, 17 eggs, 0 chickens, 500 people, or hydrogen atoms, or eclipses of the Moon, or lightning flashes. Next in order of abstraction comes the system *integers*:

$$\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$$

Initially these were defined as a convenience, introduced in order to make the laws of arithmetic more systematic and manageable. But there are many things in the world which cannot be quantified using them. We cannot, for example, accurately speak of -3 people in a room! On the other hand, there are various slightly more abstract ideas for which a description in terms of integers *is* appropriate, a bank balance being, perhaps, the example which springs most readily to mind. But in basic physics, also, the integers have their role to play. The clearest case of a physical attribute which seems to be accurately quantified by integers is electric charge. As far as can be told from accurate experiment, there is one basic unit of electric charge, namely that of a proton, and all other systems in nature have a charge which is an exact integer, positive, negative, or zero, when described in terms of this unit. (If the hypothetical quarks are, after all, eventually discovered existing as free particles – and intensive searches have so far failed to reveal them – then the values of this unit would have to be divided by three, but physical systems could again be described using integer charge values.) There are also other less familiar physical quantities which are described by integers, such as baryon number, the various leptonic numbers (presumably) and the quantum mechanical spin of a physical system about some axis.

While the arithmetical operation of subtraction is simplified by the introduction of negative integers, the operation of division is not. For that, the fractions are needed:

$$0, 1, \tfrac{1}{2}, -1, -\tfrac{1}{2}, \tfrac{1}{3}, \tfrac{2}{3}, 1\tfrac{1}{2}, 2, -\tfrac{1}{3}, \ldots$$

Any two such numbers can be added, subtracted, multiplied or divided (except by 0). But, while convenient mathematically, the system of fractions has seemed not to play any clear role

† "Raffiniert ist der Herrgott aber boshaft ist er nicht"; see, for example, Banesh Hoffman, *Albert Einstein, Creator and Rebel* p. 146, Hart-Davis, MacGibbon Ltd., London, 1972.

in quantifying features of nature. Instead, we must apparently leap further in abstraction and include all the so-called *real* numbers

$$0, 1, \tfrac{1}{2}, -1, \sqrt{2}, 2^{\sqrt{2}}, \pi, -6.1974302\ldots, \quad \ldots$$

each of which can be represented in terms of a (signed) infinite decimal expansion. The rules of finite arithmetic are precisely the same as for fractions, but the system of reals is more complete in that certain *infinite* operations can now be performed. In particular, there are such infinite sums as:

$$\frac{\pi^2}{6} = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \ldots$$

$$\sqrt{2} = 1 + \frac{1}{8}\left(\frac{2}{1}\right) + \frac{1}{8^2}\left(\frac{3\times 4}{1\times 2}\right) + \frac{1}{8^3}\left(\frac{4\times 5\times 6}{1\times 2\times 3}\right) + \frac{1}{8^4}\left(\frac{5\times 6\times 7\times 8}{1\times 2\times 3\times 4}\right) + \ldots$$

According to normally accepted views, measurements of time and distance are to be described in terms of real numbers. Indeed, it would seem that such physical notions supplied the original motivation for their invention.

The famous paradoxes of Zeno were conceived before the mathematical notion of a real number had been formally introduced. These paradoxes expressed a puzzlement about the continuous nature of time and space which had long been felt, and which have seemed to be resolved now that the rigorous mathematical concept of the real number system is at hand. In fact, such has been the success enjoyed by this number concept, that it is now hard to contemplate that time and space could be described in any other way. One gets a strong impression that time and space *are* continuous – with just the same sort of continuity that finds rigorous expression in the real number system (bearing in mind, of course, that space is three-dimensional rather than one-dimensional). But perhaps we are being blinded by appearances and by our long familiarity with the real numbers. It is this familiarity which makes *appear* natural the extrapolation to the infinitely small of a seeming continuity which is present on ordinary scales of time and space. But it could turn out that time and space do not, after all, have the kind of continuity on a very small scale that has been assigned to them. In my view – and in the views of some others[1] – such a radical change in our understanding may well be in the offing. This is a question to which I shall return later in this essay.

But if we accept the picture presented by the physics of today, the real number system provides an unending supply of elements which is called forth again and again for the quantification of physical concepts. For example, there is velocity, energy, momentum, frequency, mass, temperature, density, force, and so on. The use of real numbers for their description can, however, be traced back to the real-number continuity of space and time, so there is, in essence, nothing new to be gained from the observation that such numbers are being employed here also.

But mathematics does not stop here. Whereas subtraction, division and infinite summation have been made systematic concepts within the real numbers, the operation of solving equations has not. For example, the simple-looking equation

$$x^2 + 2 = 0$$

has no solution among the reals, whereas the apparently similar one

$$x^2 - 2 = 0$$

has the two real solutions $x = \sqrt{2}$ and $x = -\sqrt{2}$. However, for several centuries mathematicians have become used to the idea that the first equation, also, can be considered to have two solutions, written $x = \sqrt{-2}$ and $x = -\sqrt{-2}$. These are certainly not real numbers (since no negative number can have a real square root), but they are good numbers, nevertheless. Several abstractions have, in any case, had to be made in the passage from the natural numbers to the real numbers. The one further abstraction that is needed is that an extra number, denoted i, be adjoined to the system of reals, together with all those numbers formed from i by multiplying and adding real numbers to it. The number i is to be a square-root of -1:

$$i^2 = -1,$$

and the general element of the extended number system so produced, called the system of *complex numbers*, has the form

$$a + bi$$

where a and b are real numbers. In particular, the solutions to the first equation considered above are given when $a = 0$ and $b = \pm\sqrt{2}$.

At first, the notion of a complex number may seem a little bewildering. One has become accustomed to the idea that negative numbers do not "really" have square roots. It may seem, perhaps, that although these "complex numbers" are doubtless logically consistent, they serve merely as a formal device introduced for purposes of mathematical convenience. And mathematically convenient they certainly are! Not only do they enable square roots to be taken with impunity, but also cube roots, fifth roots or, indeed, any other complex power of any complex number (except perhaps zero) whatsoever. Furthermore, polynomial equations of any degree can now be solved in a completely systematic fashion. And new previously unexpected properties arise of extraordinary power and beauty. That the system of complex numbers occupies a regal place within the realm of abstract mathematical ideas, there can be little doubt. Moreover, they are not simply a "convenience"; they take on a life of their own. Yet one still may have the lingering feeling that they are not part of "reality", but merely creations of human thought.

But we must ask why it is that real numbers themselves give an impression of having this "reality" that the complex numbers seem not to have. Partly it is a question of familiarity. We are used to calculating with real numbers (or at least with finite decimal expansions) and come across them at an early stage in life. Complex numbers we encounter only much later, if at all. But more than this, it is the feeling that physical measurements, notably those of space and time, employ real numbers, whereas no such obvious physical realisation of the complex number system makes itself felt.

This, at least, seems to be the case with classical physics. With the coming of quantum mechanics the situation changes. For complex numbers play a crucial role in one of the most basic axioms of quantum theory – as will be explained shortly. But important as this fact is for the physics of the submicroscopic world, its significance begins to get lost on the macroscopic scale. Yet the role of the complex can still be discerned in the nature of phenomena at any scale

– particularly where relativity is involved. The geometry of our world is, indeed, more "complex" than it seems to be at first.

2. THE GEOMETRY OF THE COMPLEX

To try to see how this can be so, let us first indicate the standard way in which complex numbers are represented in terms of ordinary geometry. We envisage a Euclidean plane, with Cartesian axes x, y. The complex number $\zeta = x + iy$ (x, y real) is represented in this plane (called the Argand plane of ζ) by the point with coordinates (x,y). There are simple geometrical rules for the sums, products, differences and quotients of complex numbers.[2]

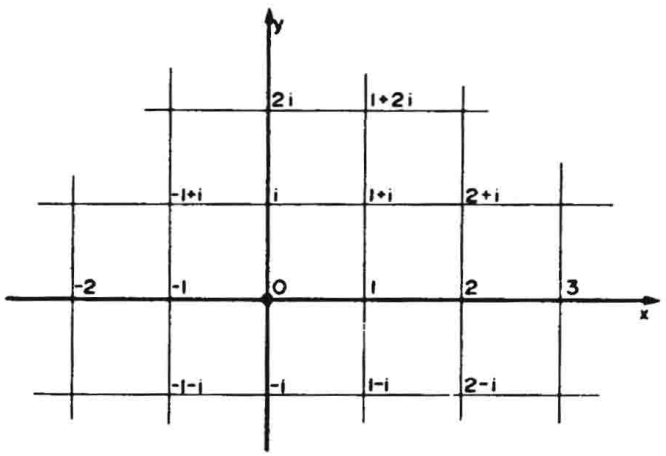


Fig. 1. The representation of complex numbers on the Argand plane.

So it seems that the world of the Flatlander could be described by the geometry of the complex number system. But what has this to do with *our* world, with its three dimensions of space and one of time? One answer is to be found in the basic rule of quantum mechanics known as *linear superposition*. If A and B represent quantum states, then for every pair of complex numbers α and β (not both zero) there exists a new state represented as $\alpha A + \beta B$, where for each distinct ratio $\alpha:\beta$ we get a physically distinct situation. Though rather formal and abstract, as stated, there is a link, apparently tenuous but in my view fundamental, between this rule and the geometry of space and time.

This may be seen at its most primitive level in the states of spin of an electron. Each such (pure) state of spin may be represented by a direction in space – where the electron is viewed as spinning in a right-handed sense about that direction. If A is the state where this direction is upwards and B where it is downwards, then any other spin direction can be represented as $\alpha A + \beta B$. There is therefore a one-to-one relation between the directions in space and the ratios $\alpha:\beta$ of pairs of complex numbers.

Now the only difference between such a ratio and just a *single* complex number is that with a ratio, *infinity* is included (i.e. when $\beta = 0$). The space of such ratios can be represented as the *Riemann sphere* S, which may be related to the Argand plane P as follows. Imagine a third axis

(the z-axis) perpendicular to the plane P and through the centre 0. Take the sphere S to have unit radius and centre at 0. The north pole N of S is the place where the positive z-axis meets S. Consider a point R of P, representing the complex number ζ. The straight line RN (extended, if necessary) meets S in a unique point Q, other than N (stereographic projection). This point Q then represents the complex number ζ, on S, where ζ is now viewed as a ratio (say $\alpha:\beta$) and where N represents $\zeta = \infty$.

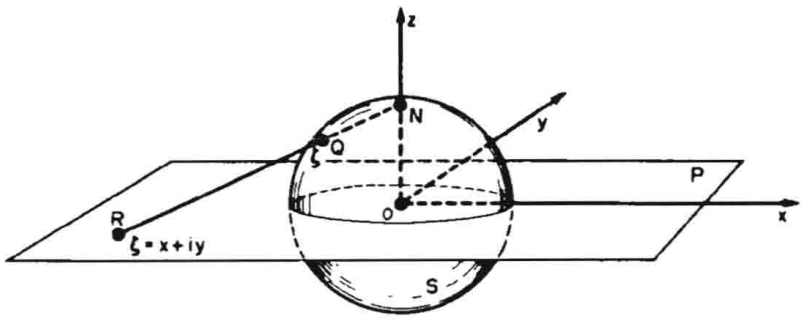


Fig. 2. Stereographic projection from the Riemann sphere to the Argand plane.

The significance of this correspondence, for our present purposes, is that if we imagine our electron situated at 0, then each pure state $\alpha A + \beta B$ of spin may be represented by a unique point of S, namely where the (oriented) spin-axis meets S. It turns out that this is precisely the point Q representing the ratio $\alpha:\beta$.

Thus, we have a remarkable tie-in between the directions in space and the fact that complex numbers are employed in the quantum superposition law. And the two-dimensionality of the complex number system is seen to be intimately related to the three-dimensionality of space.

This correspondence can be carried further. Imagine, now, an observer situated at the centre 0 of the sphere S. He looks momentarily out at the sky and represents each star that he sees by the point of S in that direction. In this way he can assign a complex number (ratio) ζ to each star as its coordinate. We now envisage a second observer in any state of motion whatever, who momentarily coincides with the first at the instant that the stars are observed. The second observer carries his own reference sphere S' and obtains a different complex number ζ' to label stars. The correspondence between ζ and ζ' is given by a remarkably simple formula:

$$\zeta' = \frac{\kappa\zeta + \lambda}{\mu\zeta + \nu},$$

where $\kappa, \lambda, \mu, \nu$ are complex parameters defining the relative motions and orientations of the two observers. This formula takes into account the fact that the two observers may be rotated relative to one another, the relativistic transformation between fast-moving observers and the finiteness of the speed of light being also correctly incorporated.

The formula is especially remarkable in that it is what is called *holomorphic*. This means that it is expressed analytically entirely in terms of $\zeta = x + iy$, the so-called *complex conjugate* $\bar{\zeta} = x - iy$ being nowhere involved. This holomorphic property is important and arises only because the problem has been treated correctly according to relativity.[3] No such result arises within Newtonian theory. The full power and beauty of complex analysis really only arises when holomorphic properties are considered. What we are beginning to see here is the first step of a

powerful correspondence between the space-time geometry of relativity and the holomorphic geometry of complex spaces.

3. TWISTORS

We have, in effect, been just studying an aspect of the geometry of the space of *photons*. Those photons which enter the eyes of our two observers are labelled holomorphically in terms of complex numbers, as a first indication of the fact that the *space of all possible photons* is really a *complex* space. That is, the different (classical) photon states can be described holomorphically by complex parameters. Photons are examples of massless particles. All such particles travel with the speed of light, but photons are distinguished from the others by the fact that they spin about their direction of motion with a magnitude $\hbar$. For the others the spin value is some different multiple of $\hbar$. Intuitively, the massless particle may be visualised in terms of a straight line drawn in space-time, namely its world-line. But this is not strictly accurate, since the particle is not quite localised when the spin is non-zero. More correctly, the state of a (classical) massless particle should be described by its energy and momentum, and its relativistic angular momentum about some origin. It turns out that all this information can be encoded as a certain complex quantity called a *twistor*[4],[5] which is (holomorphically) parameterised by four complex numbers.

These twistors may be taken as the primary elements in an approach to physics for which space-time points are superseded in their role. Instead of describing a particle in terms of a space-time locus, we describe it using twistors. We may think of a twistor as lying, conceptually, somewhere between a point and a particle. Points themselves are to be *constructed* from the twistors. And so also are the particles. In place of the wave function of a quantum mechanical particle, it turns out we require a holomorphic function of one or more twistors. The 1-twistor particles are all massless (photons, gravitons, . . .); the 2-twistor particles appear to be "leptons" (electrons, muons, . . .); and the 3-twistor particles, "hadrons" (protons, neutrons, pions, kayons, . . .) while possibly there are also 4-twistor particles, etc., forming further families (e.g. the newly discovered J/ψ particles). Moreover, the number of twistors used can apparently be increased when needed so that an n-twistor particle can also play a role as an $(n+1)$-twistor particle. The usual space-time descriptions are accordingly dispensed with altogether, although translations to space-time terms can be achieved when needed. The description of elementary particles which arises in this way must be viewed, so far, as rather speculative. Yet the mathematics seems compelling – and the classification scheme which arises appears to be compatible with what is known from standard elementary particle theory.

My contention is. then, that the twistor description of our world is more fundamental than space-time. On a classical macroscopic scale, and when general relativity is not involved, the two descriptions are *equivalent*. But for the micro-world of quantum particles (say at the level of 10^{-13} cm or less), my claim is that the twistor description will eventually give a *more* accurate picture. And the geometry of twistors is fundamentally complex, achieving an essential interweaving of quantum-mechanical principles into the ideas of geometry.

It should be emphasised, however, that twistor theory remains somewhat conjectural and incomplete at present. Although some surprisingly economical twistor expressions are already at hand for the description of much of physics, a great deal more needs to be done. The large-scale phenomena of general relativity have yet to be adequately incorporated, though

there are some very promising indications. And a twistor theory of elementary particles or of quantum fields is only just beginning to emerge. Nevertheless, the increasing role of complex spaces, as opposed to real ones, seems clear. This much arises not only from twistor theory but from many other approaches to fundamental physics. But in the twistor approach the role of these complex spaces is all-embracing. If twistor theory does indeed provide a more satisfactory framework for the description of basic physics than does the normal space-time approach, then we shall truly be able to say that nature *is* complex – in the sense of the word intended here.

4. WHERE NEXT?

What of more complicated kinds of number than the complex? There are such mathematically defined objects as quaternions, octonions and others of various different kinds. But in each case it turns out that some law of arithmetic must be sacrificed. If we wish these laws to remain intact, then complex numbers are as far as we can go. Moreover, it appears that without retaining these laws of arithmetic, a suitable function theory cannot be obtained.

It may seem ironical that complex numbers, with their own brand of two-dimensional continuity, should have had first to be brought into physics in a basic way in order to give mathematical description to various observed physical phenomena of *discreteness*. As its name implies, quantum theory is a theory designed to handle the discrete, yet it is a theory based on the use of the complex *continuum*. However, this is not so much of a paradox as it might seem at first. There is a certain kinship – even a kind of duality, reflecting the wave-particle duality of quantum physics – between complex continuity and discreteness. There is a rigidity exhibited by holomorphic functions which is quite absent in the case of functions of a real variable. Already in standard quantum theory a discreteness emerges out of the seeming morass of complex continuity. With twistor theory, it appears that we may be able to say more. In particular, the fact that electric charge occurs in integral multiples of some fixed value finds explanation within the theory – as a consequence of the rigidity of holomorphic functions.

So perhaps we shall eventually come full circle. By generalising the concept of number as far as possible away from the natural numbers of which we have direct experience, while retaining the maximum of arithmetical properties, we are apparently led back to a discreteness which we had seemed to have left behind. Perhaps, after all, the physical laws are of a discrete basically simple combinatorial nature – while at our present level of knowledge, it seems to be the complex continuum which provides the clearest route to a deeper understanding of physics.

FURTHER READING

1. E. Schrödinger, *Science and Humanism*, Cambridge University Press, 1952.
2. E. T. Copson, *An Introduction to the Theory of Functions of a Complex Variable*, Clarendon Press, Oxford, 1935.
3. R. Penrose, "The apparent shape of a relativistically moving sphere", *Proc. Camb. Phil. Soc.* **55**, 137-9 (1959).

Concerning twistor theory:

4. R. Penrose, "Twistors and particles", in *Quantum Theory and the Structures of Time and Space*, edited by L. Castell, M. Drieschner and C. F. von Weizsäcker, Carl Hanser Verlag, München, 1975.
5. R. Penrose, "Twistor theory, its aims and achievements", in *Quantum Gravity*, edited by C. J. Isham, R. Penrose and D. W. Sciama, Clarendon Press, Oxford, 1975.

93
GRAVITATIONAL COLLAPSE: A REVIEW

Physics and astrophysics of neutron stars and black holes (Proc. of the International School of Physics 'Enrico Fermi', Course LXV 1975), pp. 566–82. *North-Holland, Amsterdam*, 1978

What is perhaps a little distinctive about this one of my many introductory articles on black holes is that it refers to a crucial weakness in the commonly cited Michell/Laplace argument anticipating the possibility of entities that we indeed now refer to as 'black holes'.

214 Gravitational collapse: A review. *Physics and astrophysics of neutron stars and black holes (Proc. of the International School of Physics 'Enrico Fermi', Course LXV 1975)*, pp. 566–82

1. – Introduction.

Now that there appears to be some strong observational evidence for the existence of a black hole (in Cygnus X-1), it is, perhaps, less necessary than before to try to present convincing theoretical arguments that black holes ought to occur. Nevertheless, it should be borne in mind that, so far, the observations are of a negative character: Cygnus X-1 is argued to be a black hole simply by a process of elimination, for it is apparently compact and too massive to be either a neutron star or a white dwarf. But, since any compact object more massive than a neutron star would, in any case, be in a regime of strong gravitational fields (where gravitational potentials of order c^2 are encountered), such objects would have to be described according to a relativistic gravitational theory. So one must still rely on theoretical deductions from general relativity—or else from whatever relativistic theory one might perhaps believe to be preferable to general relativity—in order to carry the consequences of the observation to the desired conclusion: a black hole exists.

Here I shall review first the standard picture of a spherical black hole, then the arguments for believing that small deviations from spherical symmetry are unimportant and the general cosmic-censorship principle. I shall mention some properties of rotating (Kerr) black holes and, finally, something of Hawking's « black-hole explosions ».

2. – Spherically symmetrical collapse.

Let us first recall the often quoted Newtonian argument for « black holes » due to LAPLACE [1, 2]. Essentially this is as follows. The escape velocity at the surface of a (Newtonian) spherical body of mass M and radius R is $(2GM/R)^{\frac{1}{2}}$. Hence, if $R < 2MGc^{-2}$, the escape velocity will exceed c, the velocity of light, and, since light cannot escape, the object would then appear to be completely black, no matter what is its intrinsic luminosity. But, impressive and far-sighted

as now we find this prediction to be, it is not really a fair deduction within the context of Newtonian theory (*). For we must ask what speed light would be expected to have at the surface of such a body. Imagine some observer at its surface observing a distant star. The light from such a star would achieve, in addition to its natural speed c of travel from star to body, an additional component arising from the kinetic energy that it acquires in falling down to the surface of the body. (I am here adopting a particle theory of light in true Newtonian tradition.) So the speed of light at the surface should exceed the value c and, indeed, be just as to be sufficient to overcome the gravitational-potential barrier, allowing escape into space with the resulting final speed of c. (Imagine the starlight being simply reflected back into space off a horizontal mirror at the body's surface.) Of course, insisting on a unique speed for light in free space (let alone at the surface of such a body) is rather foreign to Newtonian theory. The correct discussion requires ideas of special relativity to be incorporated at the local level and cannot really be properly applied within the context of Newtonian ideas. Nevertheless, Laplace's achievement of 1795 stands out as remarkably insightful (albeit somewhat fortunate) and very far-reaching.

If we adopt, instead, a viewpoint in which special relativity is to hold locally, we are led to a picture which differs in certain essential respects from the Newtonian one. In the first place, it is clear that, if a static description of the space-time metric can be used, then light will never be trapped. For a three-dimensional spatial picture can then be adopted in which light moves outwards in all directions from every point. If light were not able to escape from a certain (closed) region X, then this region would be bounded by a « horizon » H $(= \partial H)$, with the property that light emitted from any point p of H could not enter the region complementary to X. This situation could only happen if the light emitted from p were barred from travelling in certain (spatial) directions, a possibility which has been explicitly excluded by virtue of our static description.

The concept of escape velocity still makes good sense for relativistic static space-times. If we represent the general static metric in the form

$$ds^2 = V dt^2 - \gamma_{\varrho\sigma} dx^\varrho dx^\sigma ,$$

where

$$V = V(x^1, x^2, x^3) > 0 , \qquad \gamma_{\varrho\sigma} = \gamma_{\varrho\sigma}(x^1, x^2, x^3) , \qquad \varrho, \sigma = 1, 2, 3 ,$$

and where $V \to c^2$ at spatial infinity, the quantity

$$-c(c - V^{\frac{1}{2}})$$

(*) It seems possible that LAPLACE came round to this view also, since this paper is not to be found among his final collected works.

measures the gravitational potential and

$$(c^2 - V)^{\frac{1}{2}}$$

the local escape velocity. (The justification of this will be touched upon in sect. 5.) We note that the escape velocity is always less than c, which is consistent with the above deduction that light cannot be trapped so long as a static description of the space-time can be maintained.

It would, however, be wrong to conclude that in a relativistic gravitational theory light is unlikely to be trapped. One would expect that something analogous to the Newtonian inequality $R < 2MGc^{-2}$ would still hold, as a criterion for the escape velocity to exceed c. And, since special relativity is now to hold locally, we expect that, in some appropriate sense, the value c is to remain the local light speed, even in regions where the gravitational potential is of order c^2. Thus it seems very plausible that $R < 2MGc^{-2}$, or something like it, should actually be a criterion for the trapping of light—and, consequently, that the static description of the space-time can be no longer maintained for such a situation.

Indeed, if we adopt standard general relativity as our theory of gravity, then this is actually the case. In certain other theories (for example Nordstrom's) trapping of light does not occur, so it is important, for the subsequent discussion, to decide on which theory of gravity is to be adopted. In fact I shall make no attempt, here, to consider theories other than standard general relativity. Though the tests of this theory are still not very numerous, they are by now quite impressive. All the unambiguous observations are now consistent with the theory [3]. This, together with the extraordinarily many-faceted inner consistency of the theory, strongly suggests (to me at least) that one need not look farther than general relativity in order to find an accurate mathematical description of strong gravitational fields. Indeed it is now rather hard to find a viable alternative, as the work of the Caltech group of Ni, Thorne and Will [4] has shown.

For a body with spherical symmetry we have (by Birkhoff's theorem [2, 5]) the familiar Schwarzschild metric

$$ds^2 = (c^2 - 2MG/r)\,dt^2 - (1 - 2MG/rc^2)^{-1}\,dr^2 - r^2(d\theta^2 + \sin^2\theta\, d\varphi^2)$$

holding for the region $r > R$, where R is the radius of the body (the measure of radial distance r being chosen so that the spheres $r = \text{const}$, $t = \text{const}$ have area $4\pi r^2$). Clearly, $R > 2MGc^{-2}$ for this static form of metric to hold and, by the above discussion, there can be no trapping of light, wherever this static form obtains. But Birkhoff's theorem applies to collapsing spherical bodies, so we may take R to be a function of t. (To hold R constant, for R arbitrarily close to $2MGc^{-2}$, would require an indefinitely large force to support

the surface of the body.) We allow for the possibility that R can become smaller than $2MGc^{-2}$ (*i.e.* that the body crosses its Schwarzschild radius) by transforming to a nonstatic system of co-ordinates in which, say, the co-ordinate t is replaced by the advanced time

$$v = t + rc^{-1} + 2MGc^{-3}\log(rc^2 - 2MG),$$

so that the metric takes the Eddington-Finkelstein [2, 6] form

$$ds^2 = (c^2 - 2MG/r)\,dv^2 - 2c\,dr\,dv - r^2(d\theta^2 + \sin^2\theta\,d\varphi^2).$$

For $r > 2MGc^{-2}$ (and $r > R(v)$) the space-time metric is the same as before, but now it extends to cover a larger region. We are allowed all positive values of r for this new metric form, so we can discuss the collapse of a body for which $R(v)$ can become arbitrarily small. The metric is nonsingular at the Schwarzschild surface $r = 2MGc^{-2}$ and locally the space-time is as well behaved there as anywhere else. However, the Schwarzschild surface does represent a boundary between regions where a static description obtains and where it does not.

In fig. 1, the null cones of this metric are depicted. Within the Schwarzschild surface (in the vacuum region) the null cones are all tipped over in the direction of $r = 0$. Since the world-lines of material particles must have their tangents lying within the null cones, this means that all matter in this region is inevitably forced inwards towards $r = 0$. This applies in particular to the bounding surface $r = R(v)$ of the object. Although the space-time metric will have a different form inside the body (since the vacuum equations are no longer satisfied there), that metric must smoothly fit on to the external Eddington-Finkelstein form. Thus, the null cones of the latter can be used to discuss the history of the surface of the collapsing body. As it crosses the Schwarzschild surface, no discontinuous local effects occur. However, it will be seen from fig. 1 that light can escape to external infinity only from points of the bounding surface which lie outside the Schwarzschild surface. Inside the Schwarzschild surface the light is inexorably forced in towards $r = 0$, whereas, on the Schwarzschild surface itself, it is only just possible for light to mark time always remaining at the same distance from the centre. In practice, effectively no light would do so, however. As time progresses, the radial light rays in the neighbourhood of those lying in the Schwarzschild surface diverge away from them. Any wave packet travelling along that surface would tend to disperse and become infinitely attenuated with time.

An observer a long way from the collapsing body would, in principle, still see its surface, no matter how long he waited. But, the longer he waited, the closer to the Schwarzschild surface would be the point of emission of the light that he observes. The limited amount of light that the body emits just as it

crosses this surface would have to serve for the entire infinite future of the external observer. In fact, the intensity of the observed light would die off exponentially. In a short space of time the external observer would see nothing of the collapsing body. In effect, for him, it has ceased to exist. Its gravitational field remains, however, though no ponderable body is discernible to the observer. Hence the name *black hole*.

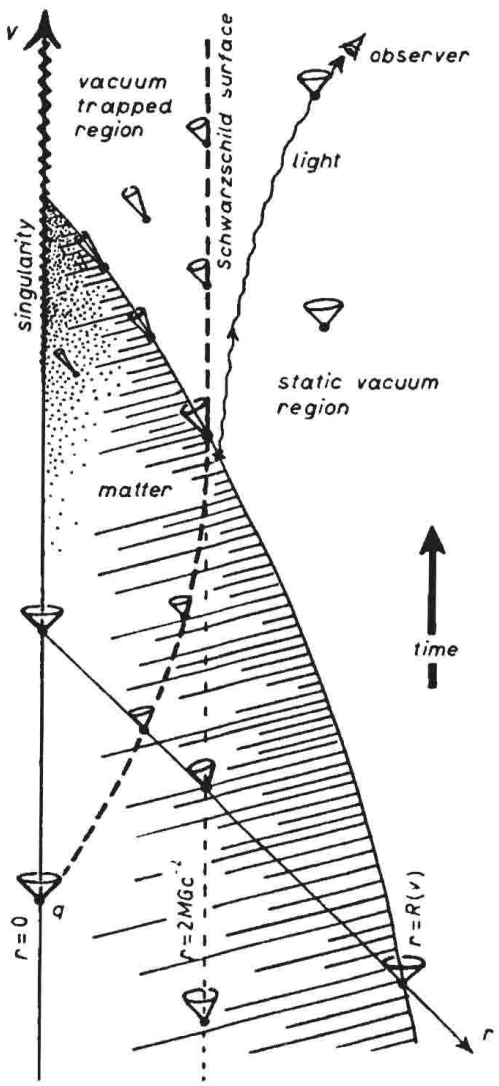


Fig. 1.

The points of the space-time may be conveniently divided into two sets. There are those from which timelike curves may be drawn into the future to infinite proper time (call this the external set) and those from which such timelike curves cannot be so drawn (the internal set). The common boundary

of these two sets is called the *absolute event horizon* (*) (or simply the horizon, in cases where there is no ambiguity with other horizon concepts). In the present situation, the horizon consists partly of the Schwarzschild surface (namely that portion of the horizon in the vacuum region) and partly of a certain null hypersurface lying inside the matter region (consisting of null geodesics emanating from a certain event q at the centre of the collapsing body and which just fit on to the Schwarzschild surface, when they leave the matter region). The external region includes the entire static vacuum region and certain events within the matter. It is the region from which particles can, in principle, escape to infinity.

Any particle which starts from an event in the internal region is inevitably forced in towards $r = 0$. This is a singularity in the space-time, in the sense that space-time curvatures become unbounded as $r = 0$ is approached. Particles can cross from the external region to the internal region, but not *vice versa*. Thus, an observer who ventured too close to the black hole would be in danger of falling into it. Once inside (*i.e.* in the internal region), there would be no escape. He would be forced towards $r = 0$. And as he approached $r = 0$, he would feel rapidly increasing space-time curvature. This curvature would manifest itself in a tidal force. His body would be stretched out in the directions towards and away from the centre, and squashed inwards in the perpendicular directions. And, as the centre is approached, these tidal effects would mount to infinity—so long as this classical space-time picture can be maintained—and the material of his body would share whatever fate was encountered by the material of the original collapsing body.

It is worth considering some orders of magnitude. Black holes are expected to form in stellar collapse, when the resulting compact object is too massive to be either a white dwarf or a neutron star. Thus, one might envisage that the smallest black hole that would be likely to arise in this way should have a mass of about $2M_\odot$. Such a black hole would have a radius of about 6 km. The tidal force for a body in free fall across the horizon would be enormous —easily enough to kill a man (the gravitational acceleration at various parts of his body differing by about a million g). At centres of globular clusters, or of galaxies, or of clusters of galaxies, there could be much larger black holes, and then the tidal force at the horizon would be much less, inversely in proportion to the square of the mass, in fact. For a black hole of $10^8\ M_\odot$ the tidal effect at the horizon would be comparable with that on a body freely falling near the Earth's surface. Thus, an observer could fall through the horizon of such a large black hole with no easily discernible effects. But then he would be forced inward and would reach the neighbourhood of the centre in something of the order of a day, during the course of which period the tidal effects would

(*) Various alternative definitions can also be given (cf. [2, 7]).

(according to our picture) mount to infinity. For a black hole of $2M_\odot$ the centre would be reached in a time of the order of milliseconds.

3. – Asymmetrical collapse.

The picture so far presented has been based on the assumption of exact spherical symmetry. We must ask whether deviations from this symmetry, small or large, are likely to change the resulting picture in any important qualitative way. There are two particular properties of the spherical-collapse picture which are especially worthy of attention. The first is the singularity at $r = 0$ and the second, the horizon at $r = 2MGc^{-2}$. At the singularity, space-time curvatures diverge to infinity, indicating that new physical laws are needed to cope with the situation at the centre. But the horizon shields these high-curvature regions from the outside world. Whatever the new physics that is needed to describe the singularity, its effects will be unobservable in the external region. But will these properties persist with asymmetries present?

Let us first ask whether the presence of a singularity is likely to be a stable phenomenon under small, but finite, perturbations of the spherical symmetry in the initial collapse. At first sight one might imagine that the spherical symmetry could be crucial. For, in that case, all the matter is falling in towards one central point, so an infinite density there is perhaps not surprising, and perhaps a feature only of the exact focussing involved. However, it turns out that this is not the case. There are various «singularity theorems» [2, 8, 9] which show that, provided the collapse is initially sufficiently close to the spherically symmetric situation, then the known classical laws imply the existence of a singularity. I shall not go into the details of such theorems here, but just outline one of the results.

Refer, first, to fig. 1 (spherical symmetry) and consider the entire future light-cone Q of a point q' which lies just to the future of q on the central axis ($r = 0$). The null hypersurface Q is generated by all the null geodesics with past end point q'. It is represented in fig. 1 by a single curve which starts at q', moves outwards and upwards through the matter region and eventually moves back in to the $r = 0$ axis in the vacuum region. The r-values start by increasing, but then they decrease again as Q collapses back to a point. For each constant-v cross-section of Q, the r-value defines its surface area as $4\pi r^2$. We have the situation of a refocussing light-cone, whereby the generators start out by diverging from one another, but then begin to come together again, finally to reconverge completely at zero surface area. Now suppose that some small but finite irregularities have been introduced into the collapsing-matter distribution. Though the situation at Q near $r = 0$ at final reconvergence might be expected to be drastically altered, not much change would be expected at the comparatively large (maximum) r-values at which the divergence of the generators of Q changes sign.

So let us take, as our indication that irreversible collapse has taken place, the existence of a light-cone which *begins to reconverge*. That is, for some point q', the divergence of the system of null geodesics through q' changes sign on each null geodesic, somewhere to the future of q'. If, in addition to this, the space-time is free of closed timelike curves, is suitably generic and is such that the energy condition holds everywhere, then the existence of a space-time singularity is implied, according to a singularity theorem of Hawking and the author [2, 9]. The energy condition used here states (with the appropriate convention for the sign of the Ricci tensor) that

$$R_{ab}t^a t^b < 0$$

for all timelike t^a. This means that, if we adopt Einstein's equations without cosmological term,

$$(T_{ab} - \tfrac{1}{2}T_c{}^c g_{ab})t^a t^b > 0$$

for all timelike t^a. In effect, this amounts to saying that in an eigentetrad of T_{ab} we have

$$\varrho c^2 + p_1 + p_2 + p_3 > 0 \quad \text{and} \quad \varrho c^2 + p_\alpha > 0 \qquad (\alpha = 1, 2, 3),$$

where ϱ is the density and p_α are the principal pressures (*).

The space-time singularity that is implied by the theorem is not necessarily of the infinite-curvature type. All that one can say is that, for some reason, the space-time manifold must be geodesically incomplete, this being what the theorem actually implies. One expects that the *reason* for the geodesic incompleteness is that the suitably maximally extended space-time manifold should possess curvatures which diverge to infinity, but only very limited exact results in this direction have been obtained so far [10].

We have, now, a sense in which the existence of a space-time singularity has been shown to be a stable phenomenon. Minor irregularities in the collapsing-matter distribution will not enable the singularity to be removed. In fact, fairly major irregularities can be allowed and the same conclusion holds true. One can fairly directly estimate the focussing effect that the collapsing matter will have on the light-cone of a point (and we recall that the focussing effect that the matter in the Sun exerts on the light which passes nearby is one of the *observed* phenomena of space-time curvature). Provided that the matter falls inwards to a region which is fairly small compared with the calculated Schwarzschild radius for the total mass involved (and this condition

(*) It we strengthen the conditions of genericity and absence of closed timelike curves to the existence of a noncompact Cauchy hypersurface, then we can weaken the energy condition to the weak energy condition $T_{ab}n^a n^b > 0$, for all null n^a [8]. This has the advantage that now a cosmological term can be included in Einstein's equations.

Gravitational collapse: A review. *Physics and astrophysics of neutron stars and black holes (Proc. of the International School of Physics 'Enrico Fermi', Course LXV 1975)*, pp. 566–82

is easily met if the mass is large enough), we expect to find a reconverging light-cone and hence deduce the existence of a space-time singularity.

The next question is the stability of the horizon. Much work has been done on this from the standpoint of conventional perturbation analysis. The conclusion is [11-13] that the Schwarzschild black-hole horizon is stable against small perturbations, where the effects of all disturbances, except those corresponding to changes in the mass and angular momentum, die away with time. In fact, angular momentum can be present in a stationary black hole. A family of rotating-black-hole solutions of the Einstein vacuum equations is known, namely the Kerr family, the different members of which are labelled by two parameters a and M, with $0 \leqslant a \leqslant M$, where MGc^{-2} is the mass, as before, and $aMGc^{-1}$ is the magnitude of the angular momentum (*). When $a = 0$, we have the Schwarzschild solution. The stability of the Kerr solutions with respect to small perturbations has also been much examined. The indications of this work [26] seem indeed to be that the Kerr black holes are all stable ($a < M$), but a complete proof is lacking as yet.

The question of the stability of the horizon of a black hole can be viewed from a different point of view. Suppose that the horizon were in fact unstable, then what could the final result of such an instability be? Two possibilities may be envisaged, namely that the configuration either evolves into some other kind of black hole (or black holes), or else that it evolves to something else. In fact we may make the discussion even more general; we need not assume that the collapse configuration ever resembled a black hole at all. The same criterion for irreversible collapse that was adopted before may be used here, namely the existence of a point whose entire future light-cone begins to reconverge somewhere to the future of that point. By the theorem quoted earlier there must be a space-time singularity somewhere. The essential question is whether this singularity is visible from infinity or not. If it *is* visible from infinity it is called a *naked singularity*. So the question is whether naked singularities can arise in gravitational collapse. If they do not, and all space-time singularities are hidden from view, then, in effect, they all lie in an «interior region» whose boundary is the absolute event horizon. The existence of such a horizon enclosing all the singularities is what (by definition) characterizes the space-time as representing a black hole. (I am not distinguishing, in this terminology, between one black hole and many black holes.) Thus, the only alternative to a black hole that we are allowed is something that is normally considered to be a more alarming possibility still, namely a naked singularity.

The conclusion that seems most reasonable to draw from the above-mentioned perturbation analysis is that, for a collapse which sufficiently closely

(*) The detailed form of the Kerr metric is given in Newman's article in this volume [14].

resembles the spherically symmetric situation, a black hole will indeed be the result. But for a collapse that is rather far from spherical symmetry, there is almost nothing that is definitely known at present. The usual presumption appears to be that no naked singularities can arise in realistic collapse. This presumption is referred to as the *cosmic-censorship hypothesis.* If cosmic censorship is true, then irreversible collapse always results in black holes.

Various attempts have been made to obtain a direct contradiction with the cosmic-censorship hypothesis [15, 16], but none has succeeded. One is left with a slight feeling that Nature is conspiring to make these attempts fail—which gives some weak support to the validity of the hypothesis. Explicit examples can be easily constructed which represent space-times with naked singularities (*e.g.*, the Weyl-Levi-Civita solutions other than the Schwarzschild solution) or which represent collapse with naked singularities [17], but for various reasons none of these can be regarded as adequately « physically realistic ». On the other hand, all attempts at a general proof of cosmic censorship appear to have met with even less success. Not even a slightly promising idea seems to have yet been suggested. The problem remains, in my view, the most important unsolved question in classical general-relativity theory.

4. – The standard collapse picture.

Suppose that an irreversible gravitational collapse takes place, in the sense that there is a refocussing light-cone Q (as in sect. 3), and suppose also that cosmic censorship holds true. The situation that then arises is the one depicted in fig. 2. There is an internal region which contains Q and also all the space-time singularities. It is bounded by the absolute event horizon H. (For this deduction we require a precise and appropriate statement of cosmic censorship [2, 18, 19]. I shall not go into the details here.) Let us suppose also that the material which subsequently falls through the horizon is limited and that the space-time finally settles down to become stationary. We now appeal to the Israel-Carter-Hawking-Robinson theorem (*) which states [18, 20-22] (modulo one remaining very minor loophole) that any stationary (vacuum) black hole must have its exterior field identical with that of a Kerr solution. It seems reasonable to suppose that the approach to stationarity is sufficiently complete for the theorem to be applied. We then infer that the final result of the collapse, irrespective of irregularities that may be present in the initial set-up, is just a Kerr black hole—a result remarkable in its completeness and simplicity, in view of the fact that the full complexity of general relativity is involved, and a totally irregular initial set-up.

(*) A full discussion of this theorem has been given by CARTER [23] in his contribution to this summer school.

We may ask what has happened to all the irregularities in the gravitational field of the original collapsing material. The answer lies in the fact that all multipole moments from quadrupole upwards can be directly changed by the emission of gravitational waves. Although the mass and angular momentum can also be changed, this is only via second-order nonlinear effects. Thus it is

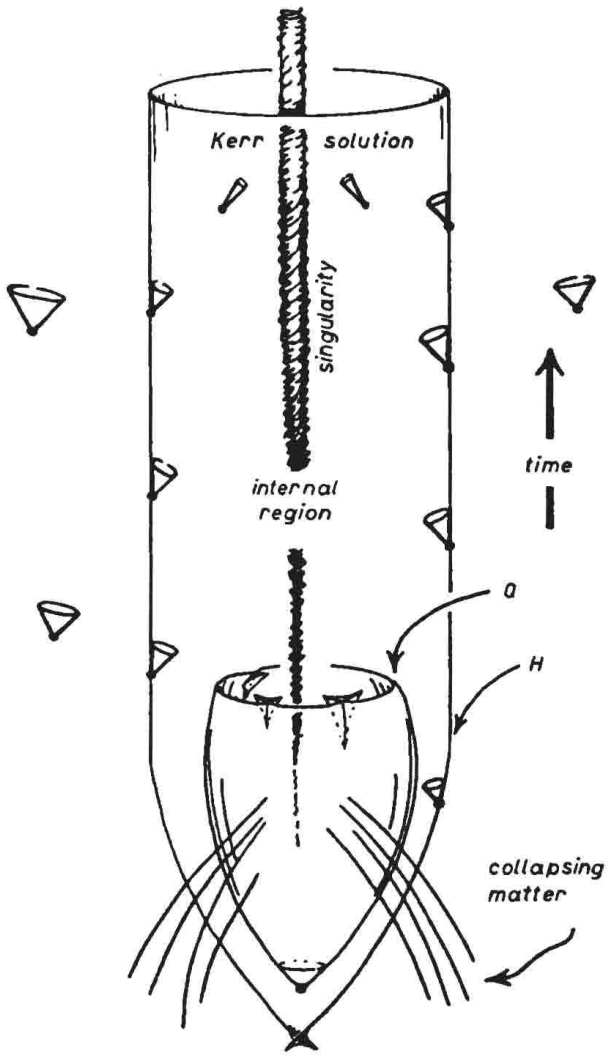


Fig. 2.

reasonable that the final mass and angular momentum should be able to take on different values (though restricted by the inequality $0<a<M$), whereas all the higher multipole moments settle to definite values, namely those given by the Kerr field. The details of the collapsing-matter distribution have rather little to do with the final field. After the matter has fallen through H, it is better to think of the black hole's continued existence and detailed structure

as more due to the self-consistency of a self-sustaining highly nonlinear gravitational field. Gravitational waves are emitted just in such a way as to adjust the quadrupole and higher moments to the required self-sustaining Kerr values.

5. – The Kerr field.

The Kerr solution [2, 24] has many remarkable properties, only a few of which will be mentioned here. These properties enable a much more complete analysis of the structure of a black hole to be carried out than one had any right to believe was possible. CARTER [25] showed how the integration of all the geodesic orbits can be reduced merely to quadratures. TEUKOLSKY [26] showed how the perturbation analysis of a black hole could be greatly simplified, and so on.

The Carter result depends upon the existence of a quadratic first integral of the geodesic equations, which arises in addition to those two linear ones which come from symmetries of the Kerr field and the quadratic one coming from the metric. These four first integrals are, respectively, of the form

$$Q_{ab}p^a p^b\,, \quad E_a p^a\,, \quad H_a p^a \quad \text{and} \quad g_{ab}p^a p^b\,(=m^2)\,,$$

where p^a is the 4-momentum tangent vector of the geodesic world-line. The constancy of each quantity arises (by virtue of $p^a \nabla_a p^b = 0$) because the relations

$$\nabla_{(a}Q_{bc)} = 0\,, \quad \nabla_{(a}E_{b)} = 0\,, \quad \nabla_{(a}H_{b)} = 0\,, \quad \nabla_{(a}g_{bc)} = 0$$

in fact hold (round brackets denoting symmetrization and ∇_a denoting covariant derivative). The middle two of these equations are examples of Killing's equations and arise because

$$\underset{E}{\pounds} \quad \text{and} \quad \underset{H}{\pounds}$$

are the operators generating metric-preserving symmetries of the space-time.

We can choose the Killing vector E^a to describe the time translation symmetry (stationarity) of the Kerr field and H^a to describe its axi-symmetry. The defining properties of E^a and H^a (up to constant factors) are that, in addition to being Killing vectors, E^a is timelike near infinity and H^a has closed orbits. Normalizing $c^{-1}E^a$ so that it becomes a unit future-pointing vector at infinity, and normalizing H^a suitably, we obtain

$$E = E_a p^a \qquad \text{and} \qquad H = H_a p^a$$

as the definitions for a test particle of conserved (test) energy and conserved

 Gravitational collapse: A review. *Physics and astrophysics of neutron stars and black holes (Proc. of the International School of Physics 'Enrico Fermi', Course LXV 1975)*, pp. 566–82

(test) angular momentum about the symmetry axis. Being linear, these two quantities are also conserved in particle collisions. The energy E must be considered to be composed both of kinetic and potential parts (to be conserved), together with the rest energy. To find the Newtonian potential of the field, one chooses p^a to be momentarily proportional to E^a so as to remove the kinetic part, and (for convenience) the test rest mass m to be unity. Then

$$E^2 = (E_a p^a)^2 = E_a E^a p_b p^b = c^2 E_a E^a ,$$

so

$$\text{Newtonian potential} = c(E_a E^a)^{\frac{1}{2}} - c^2$$

(since it is normalized to take the value zero at infinity, *i.e.* the rest energy is subtracted out). With the normal choice of time co-ordinate t we have $E^a = (\partial/\partial t)^a$, *i.e.* E^a has its t-component equal to unity and the rest zero. These facts form the basis of the statements made in sect. **2** (although generalized to the stationary case, where we need not have $V > 0$) concerning the potential and escape velocity for a static field. (Escape velocity is characterized by the fact that the total energy is equal to the rest energy, the kinetic and potential parts cancelling.)

A striking property of the Kerr field is that E^a becomes spacelike in a certain external region referred to as the *ergosphere*. In the particular (Schwarzschild) case $a = 0$, the ergosphere disappears, the external region being given precisely when E^a is timelike. Since E^a describes the time translation symmetry of the field, it is the particles whose world-lines are tangent to E^a which appear to be stationary as viewed from infinity. Thus, inside the ergosphere it is impossible for any material particle to appear to be at rest as viewed from infinity.

Outside the ergosphere, both E^a and p^a must be timelike and future-pointing. Hence $E > 0$. But inside the ergosphere there is no such restriction. By suitably choosing the timelike future-pointing vector p^a we can arrange that its scalar product with a given spacelike vector is negative. Hence we can arrange $E < 0$. That is to say, inside the ergosphere there are massive-particle orbits whose energy is *negative*!

This fact can be used as the basis of a method of extracting rotational energy from a black hole [7, 27]. A particle is thrown from infinity into the ergosphere, where it splits into two in such a way that one of the components has negative energy (and falls into the black hole), whereas the other escapes back to infinity, now possessing more mass-energy than the original particle. If we consider the level of approximation in which back-reaction of the particles on the space-time geometry is taken into account, then we would find that the negative-energy particle has slightly reduced the mass of the black hole. It would also have slightly reduced the black hole's angular momentum.

This reduction would have to be such that the quantity

$$A = 8\pi M(M + (M^2 - a^2)^{\frac{1}{2}}),$$

which measures the surface area of cross-section of the horizon H, increases slightly [27]. There is, in fact, a general result which states that, provided cosmic censorship and the (weak) energy condition both hold, then the area of cross-section of H is nondecreasing with time [2, 27].

6. – Hawking black-hole evaporation.

This principle of surface area increase, together with certain thought experiment considerations, led BEKENSTEIN [28] to suggest a relation between the surface area of a black hole and its total entropy S, given by

$$S = \frac{1}{4}\frac{kc^3}{\hbar G}A,$$

where k is Boltzman's constant. (The precise value $\frac{1}{4}$ for the numerical constant was supplied later by HAWKING [29].) According to an argument of Bardeen, Carter and Hawking [30], this association of surface area with entropy implies a corresponding association of surface gravity with temperature. This means that a nonrotating black hole of mass M must have a temperature of order

$$10^{26} M^{-1}\ \mathrm{K},$$

where M is measured in grams.

For a black hole of one solar mass ($M_\odot = 10^{33}$ g), this temperature is absurdly tiny ($\sim 10^{-7}$ K) and for larger black holes it is even smaller. Thus, if this temperature is to have any astrophysical significance, it would have to be for black holes much smaller than those formed in stellar collapse. It had been suggested by HAWKING [31] that such mini black holes might indeed be formed in the early stages of the Universe. He later showed [29] that considerations of quantum field theory on a black-hole background implied that a black hole would in fact radiate particles as though it were a black body of temperature given by the above formula. As the temperature increases the smaller the black hole is, so we have an unstable situation. Once the black hole begins to radiate more energy by the Hawking process than it absorbs from the background by accretion, it will shrink in size. And the more it shrinks, the hotter it gets, and finally it loses all its remaining mass in an explosion which momentarily creates a naked singularity (see fig. 3). A black hole of mass $< 10^{15}$ g created in the big-bang singularity (if the big bang were indeed

 Gravitational collapse: A review. *Physics and astrophysics of neutron stars and black holes (Proc. of the International School of Physics 'Enrico Fermi', Course LXV 1975)*, pp. 566–82

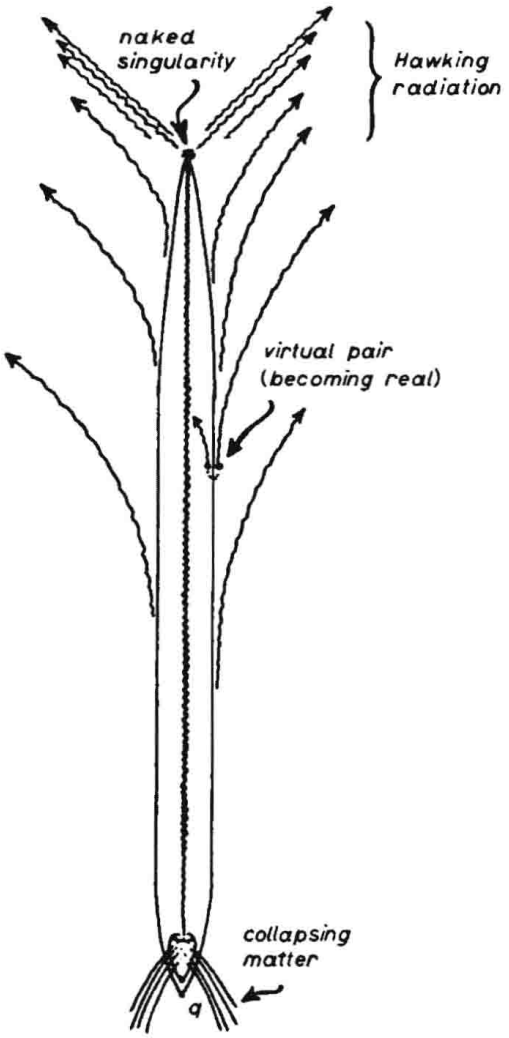


Fig. 3.

sufficiently chaotic to produce such things) would not be able to survive until the present time without exploding.

Note that this process violates both the cosmic censorship and the area principle. However, we are not now dealing with classical theory and it is possible, in effect, to violate the weak energy condition. The full discussion of the Hawking process would take us too far afield. However, HAWKING has suggested a heuristic argument which makes his process plausible. Consider a nonrotating black hole and suppose that a virtual pair of particles emerges at the neighbourhood of the horizon, one of which has positive energy and the other negative. If the negative-energy particle tunnels inside the horizon, it can arrive at a region where negative-energy orbits are achieved by real particles. It can thus become a real particle, allowing the positive-energy particle to escape to infinity and causing the black hole to reduce somewhat in mass.

In fig. 4, diagrams are drawn in which the null cones are represented at 45° to the horizontal. This makes the causality properties of the space-times easier to discern. Future and past null infinities $\mathscr{I}^+$ and $\mathscr{I}^-$ are also depicted in the diagrams. The first diagram describes the classical collapse to a black hole of fig. 1. The second describes the collapse to a black hole together with final evaporation by the Hawking process as previously depicted in fig. 3.

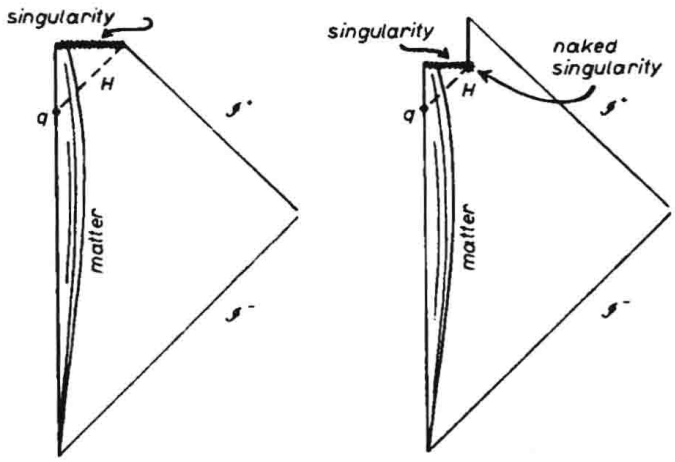


Fig. 4.

The observational relevance of the Hawking process depends upon the questionable appearance of mini black holes in the early Universe. But, irrespective of this, the process seems to be a very interesting one in view of the light that it begins to shed on some hitherto unsuspected inter-relations beetwen general relativity, thermodynamics and quantum theory.

REFERENCES

[1] P. S. LAPLACE: *Le système du monde*, Vol. **2** (Paris, 1975).
[2] S. W. HAWKING and G. F. R. ELLIS: *The Large Scale Structure of Space-Time* (Cambridge, 1973).
[3] H. A. HILL, P. D. CLAYTON, D. L. PATZ, A. W. HEALY, R. T. STEBBINS, J. R. OLESON and C. A. ZANONI: *Phys. Rev. Lett.*, **33**, 1497 (1974).
[4] K. S. THORNE, W. T. NI and C. M. WILL: in *Proceedings of the Conference on Experimental Tests of Gravitational Theories, November 11-13, 1970* (Caltech. J.P.L. Tech. Mem. 33-499, 1971).
[5] G. D. BRIKHOFF: *Relativity and Modern Physics* (Cambridge, Mass., 1923).
[6] D. FINKELSTEIN: *Phys. Rev.*, **110**, 965 (1958).
[7] R. PENROSE: *Riv. Nuovo Cimento*, **1**, Numero Speciale, 252 (1969).
[8] R. PENROSE: *Phys. Rev. Lett.*, **14**, 57 (1965).
[9] S. W. HAWKING and R. PENROSE: *Proc. Roy. Soc.*, **314** A, 529 (1970).

[10] C. J. S. CLARKE: *Gen. Rel. Grav.*, **6**, 35 (1975).
[11] T. REGGE and J. A. WHEELER: *Phys. Rev.*, **108**, 1063 (1957).
[12] C. V. VISHVESHWARA: *Phys. Rev. D*, **1**, 2870 (1970).
[13] A. G. DOROSHKEVICH, YA. B. ZEL'DOVICH and I. D. NOVIKOV: *Sov. Astron. AJ*, **11**, 233 (1967).
[14] E. T. NEWMAN: this volume, p. 554.
[15] R. PENROSE: *Ann. N. Y. Acad. Sci.*, **224**, 125 (1973).
[16] G. W. GIBBONS: *Comm. Math. Phys.*, **27**, 87 (1972).
[17] P. YODZIS, H.-J. SEIFERT and H. MULLER ZUM HAGEN: *Comm. Math. Phys.*, **34**, 135 (1973); **37**, 29 (1974).
[18] S. W. HAWKING: *Comm. Math. Phys.*, **25**, 152 (1972).
[19] R. PENROSE: in *Gravitational Radiation and Gravitational Collapse*, edited by C. DEWITT-MORETT (IAU, 1974).
[20] W. ISRAEL: *Phys. Rev.*, **164**, 1776 (1967).
[21] B. CARTER: *Phys. Rev. Lett.*, **26**, 331 (1971).
[22] D. C. ROBINSON: *Phys. Rev. Lett.*, **34**, 905 (1975).
[23] B. CARTER: preprint.
[24] P. R. KERR: *Phys. Rev. Lett.*, **11**, 237 (1963).
[25] B. CARTER: *Phys. Rev.*, **174**, 1559 (1968).
[26] S. A. TEUKOLSKY: *Astrophys. Journ.*, **185**, 635 (1973).
[27] R. M. FLOYD and R. PENROSE: *Nature, Phys. Sci.*, **229**, 177 (1971).
[28] J. D. BEKENSTEIN: *Phys. Rev. D*, **7**, 2333 (1973); **9**, 3292 (1974).
[29] S. W. HAWKING: *Comm. Math. Phys.*, **43**, 199 (1975).
[30] J. M. BARDEEN, B. CARTER and S. W. HAWKING: *Comm. Math. Phys.*, **31**, 170 (1973).
[31] S. W. HAWKING: *Mont. Not. Roy. Astr. Soc.*, **152**, 75 (1971).

94
LOCAL H^1'S AND PROPAGATION

Twistor Newsletter **7** (1978). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

This article explores the physical implications of the relationship between the (1,1) Levi-form signature of $\mathbb{PN}$ and the fact that twistor functions on $\mathbb{PT}^+$ are really H^1s. This aspect of the twistor representation of massless free fields does not seem to have been explored very much.

 Local H^1's and propagation. *Twistor Newsletter* **7** (1978). Also published in *Advances in Twistor Theory*. Ed. L. P. Hughston, Pitman, 1979

§2.12 LOCAL H^1's AND PROPAGATION *by R. Penrose*

We have now become accustomed to the fact that, at least for a region $R \subset$ that is appropriately convex (e.g. $R = CM^+$, the forward tube), *massless fr fields* of helicity s in R are represented, in twistor space, by the elemen of the *sheaf cohomology group* $H^1(Q,\mathcal{O}(-2s-2))$, where $Q \subset PT$ is the region swept out by all the lines in PT that correspond to the points of R (e.g. $Q = PT^+$ when $R = CM^+$, giving the positive frequency fields). There is an asymmetry involved in this correspondence in that Q is an extended region swept out by lines, whereas R can be a local region of CM, say a small neighbourhood of a point R of CM. However (as was pointed out to me by C.D. Hill), non-vanishing H^1's can arise also for *local* regions Q of PT. What is the space-time interpretation of these local H^1's? I shall attemp to give here a partial answer to this question.

Consider, first, a somewhat more general situation, where an n-complex-dimensional Stein manifold S is divided into two open regions S^+ and S^- by a smooth (2n-1)-real-dimensional hypersurface N whose Levi form has consta signature (p,q), with p+q = n-1. Thus, S^+ has "q degrees of convexity" at and S^- has "p degrees of convexity" at N (so that S^+ would be Stein if p = and S^- Stein if p = n-1). Assume pq > 0. Then (Andreotti & Hill 1972; Hi 1971) there is a tangential Dolbeault complex for N, whose cohomology, whi I here write $H^r(N,\mathcal{D})$ (= "$H^{0,r}$"), with r = 0,1,2,..., is given by $H^0(N,\mathcal{D}) \simeq H^0(S,\mathcal{O}), H^p(N,\mathcal{D}) \cong H^p(S^+,\mathcal{O})$, $H^q(N,\mathcal{D}) \cong H^q(S^-,\mathcal{O})$, provided that $p \neq q$; and a other $H^r(N,\mathcal{D})$'s vanish. This is assuming that we are concerned with hyper function* cohomology on N. Had we used, say, just C^∞(or even distribution

* A hyperfunction (Sato 1959, 1960) is a generalization of a distribution which is, in an appropriate sense the "dual" concept to that of a real analytic function. (Recall that a distribution is the dual concept to that of a C^∞ function.) The boundary values of holomorphic functions are alway hyperfunctions. More generally, hyperfunctions represent *jumps* between suc boundary values.

cohomology on N then the above isomorphisms would have to be replaced by inclusions. In the case p = q the only modification is that now we have $H^p(N,\mathcal{O}) \cong H^p(S^-,\mathcal{O}) \oplus H^p(S^+,\mathcal{O})$, so that, in effect, $H^p(N,\mathcal{O})$ naturally decomposes into a negative and positive "frequency" part. In each non-zero case, the space $H^r(N,\mathcal{O})$ (and $H^r(S^\pm,\mathcal{O})$) is *infinite-dimensional*. The Stein manifold S can be chosen to be an arbitrarily small (pseudo-convex) neighbourhood of a given point P of a *given* analytic real hypersurface N in a complex manifold. In this sense, these ∞-dimensional H^r's may be regarded as defined *locally* on N.

We are concerned with the case N = PN, so p = q = 1, n = 3. Thus, we have local H^1's living on PN which have, in some sense, canonically defined positive and negative "frequency" parts. What do they mean in space-time terms? A point to bear in mind is that since non-analytic cohomology (in fact hyperfunction cohomology) on N is involved, the space-time interpretation must refer to *non-analytic* behaviour in the space-time fields. My contention is that these local H^1's refer to *propagation* of massless fields. Until now, a weakness of the twistor description of fields has been that such questions (e.g. domains of dependence etc.) have been almost totally ignored. We now have the potentiality to remedy this situation. But I can, as yet, give only two partial attempts at interpretation.

The problem is, of course, that $S \subset$ PT, being Stein, can contain no projective line in PT (since a line is compact and of positive dimension). Thus the "normal" interpretation of $H^1(S^\pm,\mathcal{O}(n))$ in terms of massless fields in CM must break down. One way to get some feeling for $H^1(S^\pm,\mathcal{O}(n))$ is to "approximate" S by taking, instead, a local neighbourhood L, of a point P $\in$ PN, where $L = A \cap B$, the open sets $A,B \subset$ PT being of the "normal" type, i.e. swept out by lines in PT corresponding to the points of two disjoint sets $\hat{A}$ and $\hat{B}$ (hats

 Local H^1's and propagation. *Twistor Newsletter* **7** (1978). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

denote corresponding regions in CM). See Fig. A. Since $P \in \hat{A} \cap \hat{B}$, we have $\hat{P}$ passing through the two regions $\hat{A}$ and $\hat{B}$. Here $\hat{P}$ is an α-plane in CM, but

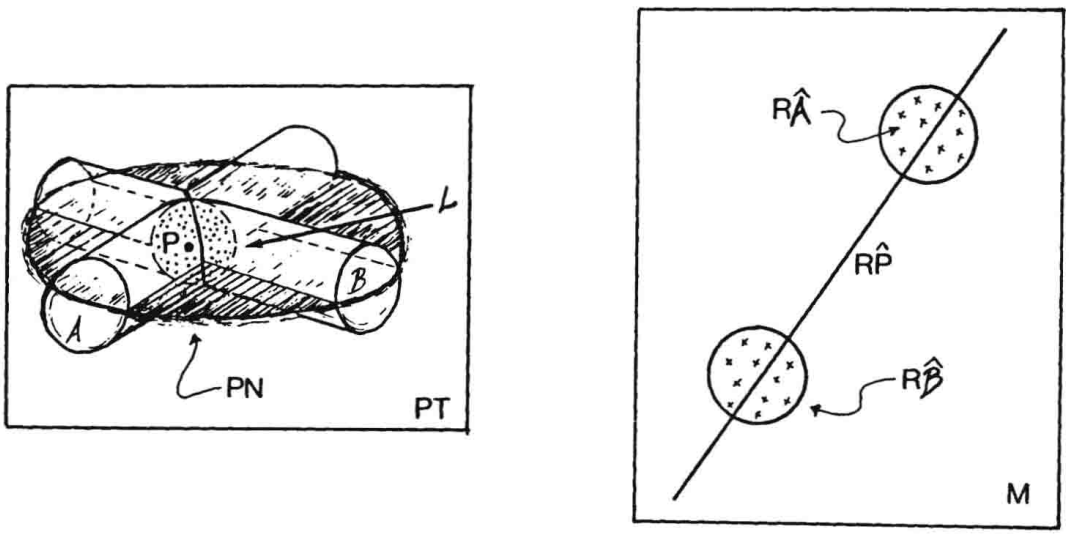


Figure A

we can equally well think in terms of $R\hat{P}$, the corresponding *null line* in M, where $\hat{A}$ and $\hat{B}$ determine subsets of M through which the null line $R\hat{P}$ passes. These two real subsets are denoted $R\hat{A}$ ($= \hat{A} \cap M$) and $R\hat{B}$ ($= \hat{B} \cap M$) (cf. Fig.A). The subset L of PN cannot be Stein since its boundary is not pseudo-convex (because the boundaries of A and B are not). But L can be *contained* in a Stein manifold S (at least if A and B are "narrow" enough, i.e. if $\hat{A}$ and $\hat{B}$ are small enough). Thus, if we can interpret $H^1(L^{\pm},\mathcal{O})$, we shall have some sort of *approximation* to $H^1(S^{\pm},\mathcal{O})$, where $L^{\pm} = L \cap PT^{+}$, provided that the contributions to $H^1(S^{\pm},\mathcal{O}(n))$ from the parts of the boundary not on PN can be ignored. By application of Mayer-Vietoris sequences etc. to this situation we find that an essential part of $H^1(L \cap PN, \mathcal{O}(-2s-2))$ describes something

like: massless free fields that exist in the union $R\hat{A} \cup R\hat{B}$ *modulo* those that can be extended to solutions of the field equations in the whole space M (the fields being hyperfunctional and of helicity s). This is an expression of the *influence* that fields in $R\hat{A}$ have on fields in $R\hat{B}$ (and vice-versa). We know, after all, that if the field in $R\hat{A}$ has a discontinuity or non-analyticity along a null geodesic (say $R\hat{P}$) that intersects $R\hat{B}$ as well as $R\hat{A}$, then this non-analyticity must show up also in $R\hat{B}$.

This seems to give an approximate interpretation of the local H^1's on PN in the neighbourhood of a point P of PN. As our sets A, B get narrower and narrower about P, so also do $R\hat{A}$ and $R\hat{B}$ get smaller and smaller. We may consider that in the limiting situation we are concerned simply with *propagation* along the null line $R\hat{P}$. Thus, $H^1(S^+) \oplus H^1(S^-)$, where S is small and Stein surrounding P, describes, in some sense, the *influence* that propagates along $R\hat{P}$.

There is another, more clear cut, way of investigating these local H^1's, as suggested to me by H. Grauert. Let S be a Stein neighbourhood of a point $P \in PN$, with $S^{\pm} = S \cap PT^{\pm}$, $N = S \cap PN$. To construct a non-zero element of $H^1(S^+)$ we find some analytic surface Σ in S whose intersection Σ^+ with PT^+ is Stein. (Pseudo-convexity at P is sufficient.) We can choose a simple situation, in fact, in which Σ is just the intersection of S with a plane, and if this is given by $A_\alpha Z^\alpha = 0$, where $A_\alpha \bar{A}^\alpha \geq 0$, then Σ^+ will indeed be Stein. Now choose a holomorphic function f (or twisted function f, a section of $\mathcal{O}(-2s-1)$) on Σ^+ which does not extend meromorphically beyond P into Σ^- (possible since Σ^+ is Stein). Take a Stein covering $\{U_i\}$ of S^+ (necessarily infinite since $\cup_i U_i = S^+$, which is not pseudo-convex) and let f_i be holomorphic in U_i with $f_i = f$ on Σ^+. Define $F_{ij} = (f_i - f_j)(A_\alpha Z^\alpha)^{-1}$. Then F_{ij} is holomorphic (section of $\mathcal{O}(-2s-2)$), since the f_i's agree on Σ^+, and the

CHAPTER 94

cocycle condition is satisfied. Now $\{F_{ij}\}$ cannot be a coboundary $F_{ij} = G_i - G_j$, because $G_i - f_i(A_\alpha Z^\alpha)^{-1}$ would then be global and meromorphic on S^+, but not extendible meromorphically beyond P - and this is impossible, since S^+ is Levi (+,-) at P.

The *twistor function* F_{ij}, having a simple pole near P, defines a *null massless free field* (Penrose 1969), taking $s \geq 1$, defined in the neighbourhood of $\hat{RP}$ in M, and which is non-analytic (possibly hyperfunctional) along $\hat{RP}$. Thus, again, we see that $H^1(S^\pm)$ has to do with the propagation of non-analytic behaviour along $\hat{RP}$.

REFERENCES

Andreotti, A. & Hill, C.D. 1972 Ann. Scuola Norm. Sup. Pisa C1 Sci. 26, 325 & 747.

Hill, C.D. 1971 In: *Partial Differential Equations*. Proc. Sympos. Pure Math., Vol. 23, p.135. Amer. Math. Soc: Providence, R.I.

Penrose, R. 1969 J. Math. Phys. 10, 38-39.

Sato, M. 1959 J. Fac. Sci. Univ. Tokyo, Sect. I 8, 139-193.

Sato, M. 1960 J. Fac. Sci. Univ. Tokyo, Sect. I 8, 387-437.

95
EXTENDED REGGE TRAJECTORIES

J. Phys. Math. Gen. **11** (1978). With G. A. J. Sparling and Tsou Sheung Tsun

It seems to me that there is something unexplained (as far as I am aware) in the observations of the straight Regge trajectories of strongly interacting particles, where pairs of trajectories join up under the joint replacements $m^2 \mapsto -m^2$, $j \mapsto -j-1$. This observation was stimulated by considerations of twistor theory, but does not appear to be (yet) explained by it. See also Chapter 192.

J. Phys. A: Math. Gen., Vol. 11, No. 9, 1978. Printed in Great Britain.

LETTER TO THE EDITOR

Extended Regge trajectories

Roger Penrose†, George A J Sparling‡ and Tsou Sheung Tsun†

†Mathematical Institute, Oxford, UK
‡Department of Physics and Astronomy, University of Pittsburgh, Pa, USA

Received 24 July 1978

Abstract. It is proposed that in the classification of elementary particles (with mass) a quantum number j plays a role, where $j+\frac{1}{2}=\epsilon(J+\frac{1}{2})$, with $\epsilon=\pm1$, J being the normal spin. This systematises the data concerning Regge trajectories, pairs of trajectories now uniting into single straight extended trajectories, where j is plotted against ϵm^2. For baryons, ϵ is found to be the 'naturality' of the parity, while for mesons, passing from $\epsilon=1$ to $\epsilon=-1$ leaves the parity unaltered.

In this note we point out a certain rather remarkable but simple property that appears to be exhibited by the data concerning Regge trajectories, and which does not seem to have been noted before. We are also proposing a somewhat unconventional way of interpreting this observation, although our primary concern, here, is not with theory (which we hope to discuss elsewhere) but with the data itself.

In conventional terms, our observation can be stated simply in the following way. Using a standard Chew–Frautschi plot, with J (the spin) plotted against m^2, we find very closely, that for the baryons, for each of the cases N, Δ Λ, Σ, Ξ separately, the leading natural parity trajectory is not only straight and parallel to the leading unnatural parity trajectory, but is such that *the two trajectories* (nearly coincident in the case of N and Λ) *intersect the J-axis in two points whose mid-point is* $J=-\frac{1}{2}$. In the case of the mesons, the statement is essentially the same but where now both trajectories must have the *same* naturality and we refer, instead, to the leading trajectory and a daughter.

This general property of intercepts appears to have no obvious explanation from standard Regge theory, but we had anticipated that something of the sort should hold on the basis of the unconventional proposal that we are putting forward. This, in effect, is that in the classification of hadron states there is a (new) quantum number j, which can take both positive and *negative* integral or half-integral values and for which $j(j+1)=\vec{J}^2=J(J+1)$(so that $j+\frac{1}{2}=\pm(J+\frac{1}{2})$, units with $\hbar=1$), J being the usual (non-negative) spin quantum number. This follows an earlier analogous proposal (Sparling 1977) which relates, more particularly, to the classification of leptons. The idea springs partly from a mathematical result (Bott 1957) concerning the representations of compact Lie groups: two such representations may be, in a sense, *different* (e.g., here, the j and $-j-1$ representations for the rotation group) even though they are, in the conventional sense, equivalent. We had been initially led to consider this idea while pursuing the implications of a twistor description of elementary particles (Perjés and Sparling 1976, Perjés 1975, 1977, Penrose 1975, 1977) but the idea has wide implications that are quite independent of twistor theory. Here we point out how the proposal fits in with the

experimental data concerning Regge trajectories in a surprising way. The above-mentioned observation takes the following form: each leading trajectory joins very accurately on to a second trajectory (or to a different parity part of the 'same' line, in the case of N and Λ), plotted in the reverse direction, where the value of $j+\frac{1}{2}$ is deemed to have opposite signs for the two portions of what we now consider to be, in each case a *single (remarkably straight) extended Regge trajectory.*

To exhibit this feature we require a slight modification of the usual Chew–Frautschi plot (Collins (1977) and bibliography therein, Chan Hong-Mo and Tsou Sheung Tsun (1978) for an up-to-date review), where now j is plotted against ϵm^2, with $\epsilon = \mathrm{sgn}(j+\frac{1}{2})$, so that $j+\frac{1}{2} = \epsilon(J+\frac{1}{2})$. Introduce ν to represent the 'naturality' of the parity, defined according to $\nu = P(-1)^{J-\frac{1}{2}B}$ (where P denotes the parity and B the baryon number). We find that $\epsilon = \nu$ in the case of baryons, yielding a striking straightness for the extended trajectories in each case (figure 1). For the mesons, all particles on the same extended trajectory turn out to have the *same* naturality ν. Thus we have no independent means of determining ϵ (and consequently j) for mesons, but find that ϵ can be assigned opposite values for the leading trajectory and one of the 'daughter' trajectories, the two combining to give a remarkably straight extended trajectory in each of the five cases where well-established 'daughters' are seen (figure 2).

One significant consequence of these plots, if our point of view is accepted, is that they can provide a good interpolated determination of the value for the intercept with $m^2 = 0$, this being theoretically important for the high-energy behaviour of scattering cross sections (Collins 1977 and Chan Hong-Mo and Tsou Sheung Tsun 1978). We should emphasise, however, that we do not necessarily anticipate that all interpolated points at the central part of the extended trajectory should be occupied (and we point out that $j = -\frac{1}{2}$ must always be *un*occupied since the corresponding number spin states is $|2j+1| = 0$). We make no suggestions as to what 'nonsense' mechanisms might be operative in our scheme to suppress resonances generally.

Since the role played by daughter trajectories now seems different from what it had been in the conventional viewpoint we may ask how other possible daughter trajectories are to be treated. We would have difficulty in fitting in many parallel daughters because of the Froissart bound, according to which the extended trajectory's intercept with $m^2 = 0$ should presumably now lie between $j = -2$ and $j = +1$. But points of a lattice can always be joined up in different ways, and we are led to entertain a tentative suggestion that there might be extended daughter trajectories with shallower slopes from those of the leading extended trajectory. In fact we find, in some cases, that straight lines can indeed be drawn through the *ground state point* to line up several daughter points into such shallower extended trajectories. There may well be other ways of interpreting, within our scheme, the various scattered points that do not lie on the main trajectories.

We make no specific proposal concerning the new families of particles (Dalitz *et al* 1977, Herb *et al* 1977), except to say that we would anticipate j playing a role here, also.

Finally, we point out that there are many predictive ways that our idea can be used. In particular, we can make some clear-cut predictions concerning the spin-parity assignments to resonances. For the N^* and Λ, since the conventional natural and unnatural partity trajectories show degeneracy with an intercept around $-\frac{1}{2}$, our predictions are the same as those of ordinary linearly rising Regge trajectories. For the ot er trajectories, however, our predictions on spin-parity will be non-trivial. For instance, we would assign $J^P = \frac{15^+}{2}, \frac{19^+}{2}$ to the $\Delta(2850)$ and $\Delta(3230)$ respectively. As another example, we would predict the existence of an $I = \frac{1}{2}$ (unnatural parity) straight

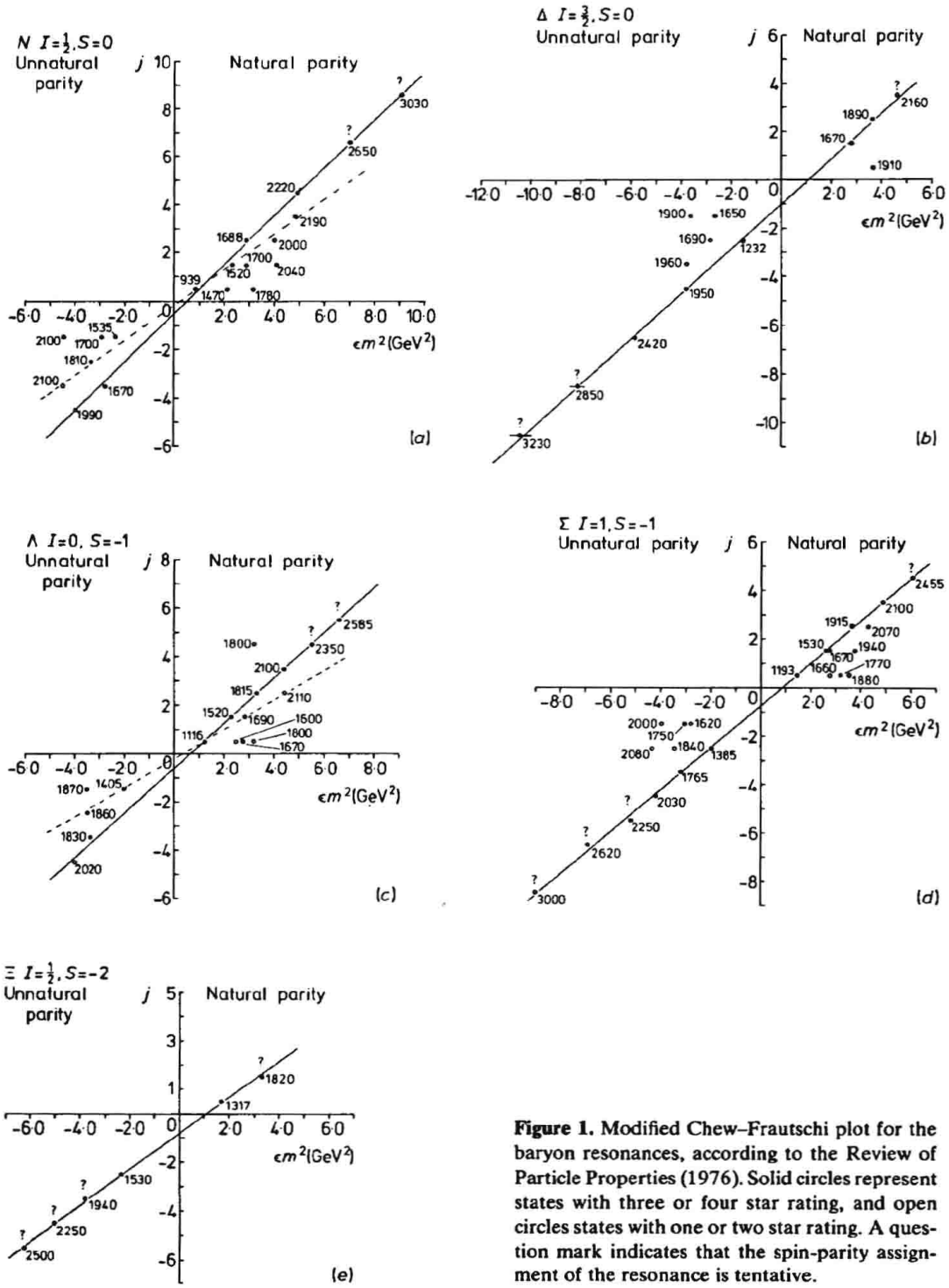


Figure 1. Modified Chew–Frautschi plot for the baryon resonances, according to the Review of Particle Properties (1976). Solid circles represent states with three or four star rating, and open circles states with one or two star rating. A question mark indicates that the spin-parity assignment of the resonance is tentative.

Regge trajectory corresponding to the extension into the negative region of the line of figure 2(*f*). This would mean, in the ordinary way of plotting things, a 'daughter' line parallel to *KL* and displaced downwards by a *J* value of approximately $\frac{1}{2}$.

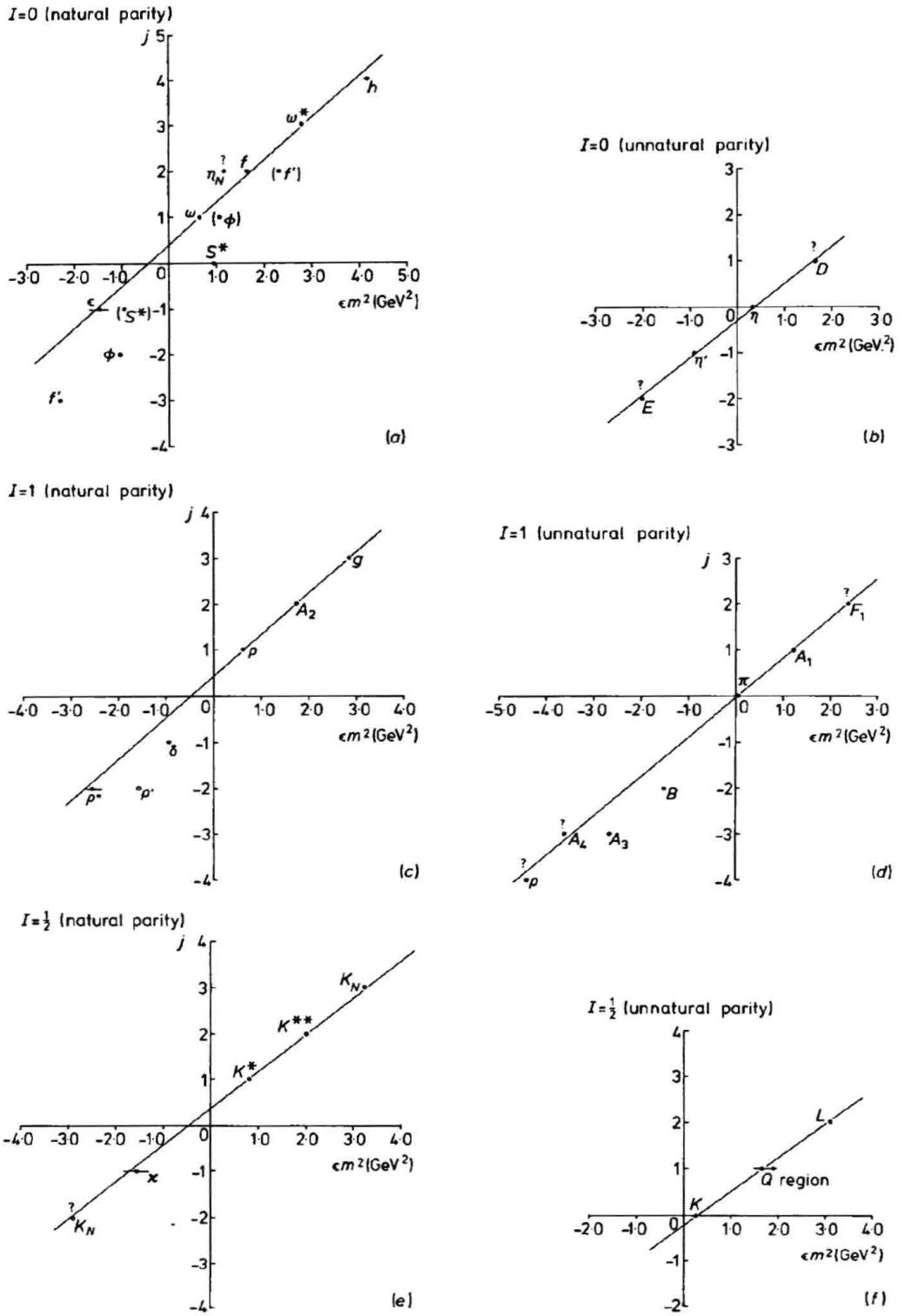


Figure 2. Modified Chew–Frautschi plot for the meson resonances, according to the Review of Particle Properties (1976). Solid circles represent states included in the Meson Table, and open circles left out from it. A question mark indicates that the spin-parity assignment of the resonance is tentative. In (*a*) brackets are put around alternative positions of resonances already marked elsewhere corresponding to a different value of ϵ.

We are most grateful to L P Hughston for valuable discussions.

This work was partially supported by the Science Research Council under Contract Number GR/A/05498.

References

Bott R 1957 *Ann. Math.* **66** 203

Chan Hong-Mo and Tsou Sheung Tsun 1978 *Phys. Rep.* to be published

Collins P D B 1977 *An Introduction to Regge Theory and High Energy Physics* (Cambridge: Cambridge University Press)

Dalitz R H, Richter B, Wiik B and Toner W T 1977 *Proc. R. Soc.* A **355** 411

Herb S W *et al* 1977 *Phys. Rev. Lett.* **39** 252

Penrose R 1975 *Quantum Theory and the Structures of Time and Space* eds L Castell, M Drieschner and C F von Weiszacker (Munchen:Hauser)

—— 1977 *Rep. Math. Phys.* **12** 65

Perjés Z 1975 *Phys. Rev.* D **11** 2031

—— 1977 *Rep. Math. Phys.* **12** 193

Perjés Z and Sparling G A J 1976 *Central Research Institute for Physics, Budapest and University of Pittsburgh Preprint*

'Review of Particle Properties' 1976 *Rev. Mod. Phys.* **48** no. 2 part II S1

Sparling G A J 1977 *University of Pittsburgh Preprint*

96
SINGULARITIES OF SPACETIME

Theoretical principles in astrophysics and relativity. Eds N. R. Lebovitz, W. E. Reid, and P. O. Vandervoort. Chicago University Press, 1978

Space-time singularity structure is investigated using TIPs and TIFs (see Chapter 50), and various forms of cosmic censorship (Chapters 40 and 271) are proposed. In Figure 7 on p. [236], the situation that Stephen Hawking would later refer to as a 'thunderbolt' is depicted. The Weyl Curvature hypothesis is referred to in the final paragraph (see also Chapters 86 and 91).

244 Singularities of spacetime. *Theoretical principles in astrophysics and relativity.* Eds N. R. Lebovitz, W. E. Reid, and P. O. Vandervoort. Chicago University Press, 1978

9 Singularities of Spacetime

Roger Penrose

1. Introduction

Combined evidence from observation and theory points to the existence of singularities in the spacetime continuum of our universe: in the first place, there is the big-bang singularity representing its initial state; second, there are the singularities inside black holes, representing the final state of collapse of massive stars. Other types of spacetime singularities might also exist. I shall review here our present understanding of the status of these singularities without, however, attempting to be comprehensive. This account will, to some extent, represent my own personal view. I am glad to be able to present this account as an article in Chandra's honor—particularly in view of the fact that it was his work of 40 years ago which led us, apparently inescapably, into the perplexities and singularities of the black-hole phenomenon.

Let us first ask what is meant by the phrase "spacetime singularity." In fact it has rather a spectrum of meanings, depending upon the context in which it is used. When speaking descriptively or in loose physical terms, we refer to singularities as regions of infinite spacetime curvature. Physically, spacetime curvature finds its expression in the gravitational tidal forces exerted on freely moving bodies. Thus, if a singularity is approached, we expect these tidal forces to mount to infinity. Ordinary physical objects could not survive too close to such a singularity. The tidal forces there would be expected to become so strong that molecules, atoms, and nuclei would in their turn be ripped to pieces; and there is no reason to suppose that elementary particles themselves would not be subject to the same fate. Tidal forces may be envisaged so large that they correspond to radii of spacetime curvature as little as the characteristic strong-interaction size: 10^{-13} cm. And it would seem that precious little of existing elementary-particle physics could survive for a background spacetime as curved as the elementary particles themselves. Furthermore, without particles with which to probe the geometry, it is hard to see how the description of a continuous background spacetime could then be meaningfully maintained.

This notwithstanding, it appears that the majority of physicists who actively concern themselves with quantization of (and in) gravitational fields believe that a drastic revision of physical theory is not required at this stage. Instead,

Roger Penrose is at the Mathematical Institute, Oxford, England.

Singularities of spacetime. *Theoretical principles in astrophysics and relativity.* Eds N. R. Lebovitz, W. E. Reid, and P. O. Vandervoort. Chicago University Press, 1978

they argue, the classical notions of spacetime geometry must hold good until radii of spacetime curvature are encountered of the order of the Planck length: 10^{-33} cm. But certainly at such more extreme situations there seems no question that new physics would be required. For this is the scale at which quantum fluctuations of the spacetime metric would be large enough to render the normal classical picture of a (pseudo-) Riemannian manifold completely inadequate as a picture of reality. Even if the required theory took merely the form of a successful application of standard quantization techniques to Einstein's equations (which, in my view, is most unlikely), this would nevertheless still have to qualify as "new" physics in view of the impasse to which such attempts seem so far to have led (see Isham *et al.* 1975).

Thus, the *physical* interpretation of the term "spacetime singularity" is a region in which curvatures have become so enormously large that our normal physical theories break down. (Whether one expects this to occur for radii of 10^{-13} cm, 10^{-33} cm, or some other value is a question to which I shall return later.) But from the point of view of a precise *mathematical* treatment, such an interpretation is rather negative and unmanageable. In the absence of a good physical theory of such high-curvature regions, we must concentrate, instead, on those regions where the normal pseudo-Riemannian picture of spacetime still holds good. We need assume not that the accuracy of our mathematical model is perfect, but that it holds good to a high degree down to the kind of dimension (say 10^{-13} cm or 10^{-33} cm) that we are concerned with. For radii of curvature smaller than that dimension, our model would be meaningless as a picture of reality. If desired, those portions of high curvature can be deleted from the model, leaving a pseudo-Riemannian manifold which is mathematically extendable to a somewhat larger one (namely, to the one obtained by reinstating the deleted portions). However, normally it is mathematically more convenient to include the highly curved portions and to work instead with a maximally extended manifold. In any case, the singularities themselves are not to be part of the pseudo-Riemannian manifold (which is viewed as consisting of mathematically "nonsingular" points only). Instead, these singularities constitute a kind of boundary (or portion of a boundary) for the spacetime manifold. (Considered in this way, the singularities for specific models need not always actually correspond to regions of infinite curvature.) The boundary can be constructed according to any of several precise (though unfortunately inequivalent) mathematical prescriptions. I shall describe only one of these prescriptions here (in § 2), namely, that of ideal points (or causal boundary), since it is easy to apply and seems closest to the physics; and a number of familiar examples will be discussed.

In § 3 I shall briefly review the reasons for believing that gravitational collapse, with or without symmetries, can lead inevitably to the existence of spacetime singularities; and I shall describe the standard picture of gravitational

 Singularities of spacetime. *Theoretical principles in astrophysics and relativity.* Eds N. R. Lebovitz, W. E. Reid, and P. O. Vandervoort. Chicago University Press, 1978

collapse to a black hole. When large deviations from spherical symmetry are present, the validity of the standard picture requires the hypothesis of cosmic censorship, according to which naked singularities are forbidden from arising in a collapse. I shall mention various versions of this hypothesis in § 4 and relate it to such questions as global hyperbolicity (the existence of a Cauchy hypersurface) and the principle of surface-area increase for black holes' event horizons.

A recent development in the theory of black holes is Hawking's quantum-mechanical evaporation process. This can have observational significance only if very small black holes (of radius down to 10^{-13} cm or less) can have been produced in a sufficiently chaotic initial big bang. But the process has in any case a considerable theoretical interest. I shall discuss (in § 5) some aspects of this and present some views of my own (which differ in certain essential respects from those of Hawking) as to the implications for thermodynamic time-asymmetry and its relation to the detailed structure of spacetime singularities.

2. TIPs and TIFs

There are various different ways that boundary points, representing singularities, may be adjoined to an (inextendable) spacetime manifold (Geroch 1968; Seifert 1971; Schmidt 1971, 1972; Hawking and Ellis 1973). The construction that has been regarded as the most mathematically satisfactory is the b-boundary of Schmidt (and its conformal extension the c-boundary; see Schmidt 1974). However, his procedure is rather nonintuitive and is difficult to apply in practice. Furthermore, according to a recent result of Bosshard (1975) it can sometimes produce anomalous results—even when applied to so apparently straightforward an example as a closed Friedmann universe. For such reasons I shall adopt here the much simpler notion of causal boundary (Geroch *et al.* 1972; Hawking and Ellis 1973). Though it may lack the richness of structure possessed by the b-boundary, and though there is the awkwardness that the past and future boundary points cannot easily be fitted together into a single scheme, there is a basic simplicity about the construction which makes the idea physically direct and easy to apply.

I shall suppose that the spacetime $\mathfrak{M}$ is strongly causal (although past- and future-distinguishing [Geroch *et al.* 1972] would suffice). Then any point $p \in \mathfrak{M}$ is uniquely defined by the set $I^+(p)$ ($\subset \mathfrak{M}$) of points lying to the chronological future of p. Similarly, p is uniquely defined by its chronological past $I^-(p)$. Writing $x \ll y$ for the statement "a future-directed timelike curve from x to y exists in $\mathfrak{M}$," where $x, y \in \mathfrak{M}$, we have (Penrose 1972) $I^+(p) := \{x : p \ll x\}$ and $I^-(p) := \{y : y \ll p\}$. Also, for any subset S of $\mathfrak{M}$ we use the notation

$$I^+[S] := \bigcup_{p \in S} I^+(p) \quad \text{and} \quad I^-[S] = \bigcup_{p \in S} I^-(p) .$$

Thus each point p may be represented either by the set $I^+(p)$ or by the set $I^-(p)$.

The advantage of representing p in this way is that a slight generalization of these sets enables us to represent not just ordinary points p of $\mathfrak{M}$ but certain extra "ideal points" which may be thought of as constituting a kind of boundary to $\mathfrak{M}$. This generalization rests on the fact that there are two intrinsic properties of the sets $I^+(p)$ and $I^-(p)$ which are almost, but not quite, sufficient to characterize them as sets of this kind; and the extra sets which arise give the ideal points.

In the first place, each set $I^+(p)$ is a *future-set* and each $I^-(p)$ is a *past-set*. (A future-set is a subset F which is its own chronological future: $F = I^+[F]$; a past-set P is its own chronological past: $P = I^-[P]$.) In the second place, each of $I^+(p)$ and $I^-(p)$ is actually *irreducible* in the sense that $I^+(p)$ cannot be expressed as the union of two future-sets each distinct from $I^+(p)$, and similarly for $I^-(p)$. An irreducible future-set is called an IF; and an irreducible past-set an IP. Those IFs which are of the form $I^+(p)$ for some $p \in \mathfrak{M}$ are called PIFs (proper IFs), while those *not* of this form are called TIFs (terminal IFs). Similarly, the IPs are subdivided into PIPs, having the form $I^-(p)$, and TIPs. By remarks above, the PIFs are in a natural one-to-one correspondence with the points of $\mathfrak{M}$; and, similarly, so are the PIPs. But the TIFs give us something extra, and so also do the TIPs. These correspond to the ideal points for $\mathfrak{M}$.

Roughly speaking, we may regard each TIF as a set "$I^+(q)$," where q is not a point of $\mathfrak{M}$ but an ideal *past boundary point* of $\mathfrak{M}$; similarly, each TIP may be regarded as "$I^-(r)$," where r is an ideal *future boundary point* of $\mathfrak{M}$. (For some spacetime models it seems natural to identify certain past boundary points with certain corresponding future boundary points, but I shall not enter into this here.) Thus, we think of the TIF itself as the set of points of $\mathfrak{M}$ lying to the (chronological) future of the ideal point which the TIF serves to define. Similarly, each TIP is the chronological past of the ideal point it defines.

There is another somewhat more intuitive way of discussing IPs and IFs which arises because of the following theorem (Geroch *et al.* 1972):

Theorem. *The IFs of $\mathfrak{M}$ are precisely the sets of the form $I^+[\gamma]$; and the IPs of $\mathfrak{M}$, those of the form $I^-[\gamma]$, where in each case γ is some timelike curve in $\mathfrak{M}$. The same holds if γ is allowed to be null.*

If γ has a past endpoint p, then clearly $I^+[\gamma] = I^+(p)$ and we have a PIF. The TIFs arise when γ is *past-endless* in $\mathfrak{M}$ (i.e., it continues indefinitely, in $\mathfrak{M}$, into the past, perhaps approaching a singularity). Correspondingly, $I^-[\gamma]$ is a PIP or TIP according as γ has a future endpoint or is *future-endless* in $\mathfrak{M}$. Thus, we may think of the past boundary of $\mathfrak{M}$ as supplying ideal past endpoints to timelike (or null) curves which are past-endless in $\mathfrak{M}$ and the future boundary as correspondingly supplying ideal future endpoints to future-endless timelike (or null) curves. For two such curves γ and γ', the condition for their past end-

points, whether ideal or not, to coincide is $I^+[\gamma] = I^+[\gamma']$; similarly the condition for their future endpoints, ideal or not, to coincide is $I^-[\gamma] = I^-[\gamma']$.

The ideal points thus constructed need not be singular points for $\mathfrak{M}$, but may be points at infinity. As an example, let $\mathfrak{M}$ be Minkowski space. If γ and γ' are (endless) timelike straight lines in $\mathfrak{M}$, then clearly $I^+[\gamma] = I^+[\gamma'] = \mathfrak{M}$ and $I^-[\gamma] = I^-[\gamma'] = \mathfrak{M}$, so γ and γ' share the same future ideal boundary point (called i^+) and also the same past ideal boundary point (called i^-). But there

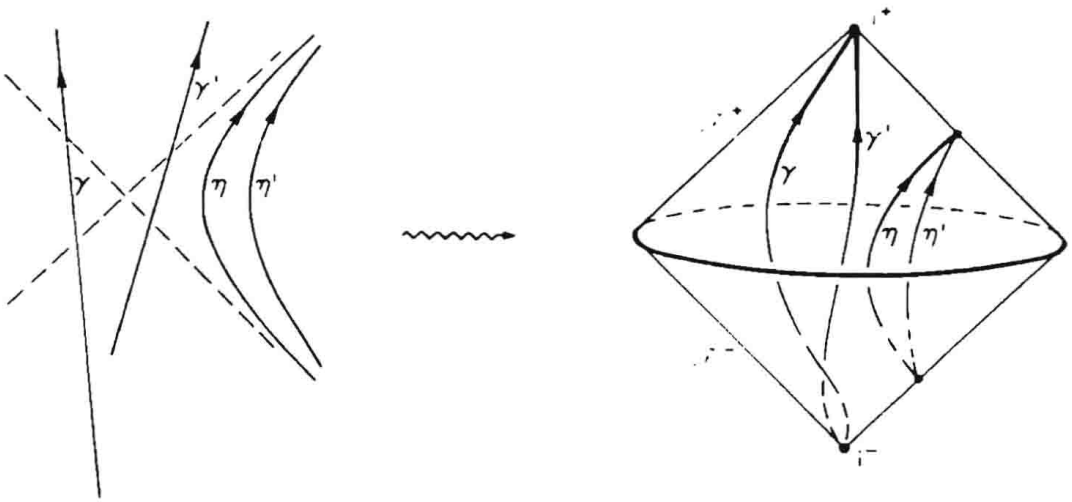


Figure 1. Timelike straight lines γ, γ' in Minkowski space acquire the ideal endpoints i^+ and i^-. The curves η, η' of uniform acceleration acquire ideal endpoints at null infinity $\mathcal{I}^+, \mathcal{I}^-$.

are also other boundary points for $\mathfrak{M}$. Consider the curve η of uniform acceleration, given in the usual coordinates by

$$x = A\,, \quad y = B\,, \quad z = (C + t^2)^{1/2}\,, \quad t \in (-\infty, \infty)\,,$$

where A, B, and C are constants, with $C > 0$. Then the set $I^+[\eta]$ consists of the points of $\mathfrak{M}$ for which $t > -z$; and the set $I^-[\eta]$, the points for which $t < z$. Thus, whatever the values of A, B, and C, the curve η acquires the same pair of ideal points at both past and future infinity. If the z-equation is replaced by $z = t$, then we get a null straight line with the same future ideal point as before, but a different past ideal point (corresponding to the TIF given by $t > z$); if it is replaced by the null line $z = -t$, then the past ideal endpoint is the same as before, but not the future endpoint. If the z-equation is replaced by $z - D = [C + (t - D)^2]^{1/2}$, then the future ideal point is the same as before, but the past ideal point depends on D (with TIF given by $t + z > 2D$); if by $z + D = [C + (t - D)^2]^{1/2}$, then it is the future endpoint that depends on D. We can obtain other ideal points by simply performing a rotation in x, y, z to the above construction, the general TIF (other than $\mathfrak{M}$ itself) being given by

$$t > lx + my + nz + p\,,$$

and the general TIP (other than $\mathfrak{M}$) by

$$t < lx + my + nz + p\,,$$

where $l^2 + m^2 + n^2 = 1$. The corresponding past and future ideal boundary points for $\mathfrak{M}$ constitute two three-dimensional systems (parametrized by l, m, n, p), each with topology $S^2 \times R$, called $\mathcal{I}^-$ (past null infinity) and $\mathcal{I}^+$ (future null infinity), respectively (see Figure 1).

As a second example, consider a Friedmann cosmological model with an initial "big bang" singularity. It is well known that such models possess *particle horizons*. Thus, if γ and γ' are the world-lines of two different substratum particles, then no point of γ' is "visible" from points of γ within a certain distance of the initial singularity; in other words, $I^+[\gamma']$ does not contain the whole of γ. It follows that $I^+[\gamma'] \neq I^+[\gamma]$, so γ' and γ have different ideal past endpoints. In fact, there is precisely one past boundary point corresponding to each sub-

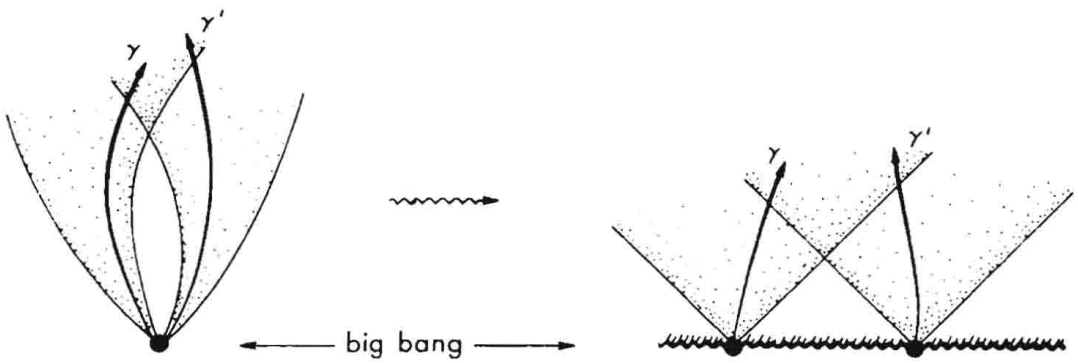


Figure 2. The TIFs representing the big bang define a three-dimensional system of singular ideal points.

stratum world-line. The initial singularity may be identified with this entire past boundary. It is spacelike (in a well defined sense to be described shortly) and has topology S^3 (if $k = +1$) or R^3 (if $k = 0$ or -1). In a similar way, we may consider future boundary points and associate them with *event horizons*. (The boundary of $I^-[\gamma]$ is the event horizon of the observer whose world-line is γ.) In a Friedmann universe model which recollapses to a final singularity, the structure of the future boundary is similar to that of the past boundary and both boundaries refer to singular points. But in a Friedmann model which expands indefinitely to infinity, the future boundary may differ considerably in structure from the past boundary. This is over and above the fact that the past boundary represents a singularity whereas the future boundary points are points at infinity.

In depicting the spacetimes with their boundaries attached, it is often convenient to use a two-dimensional representation in which null directions are

 Singularities of spacetime. *Theoretical principles in astrophysics and relativity.* Eds N. R. Lebovitz, W. E. Reid, and P. O. Vandervoort. Chicago University Press, 1978

oriented at $\pm 45°$. This entails that the big-bang singularity is depicted not as a single point but as a spacelike line (cf. Fig. 2). In Figure 3 a number of examples of familiar spacetimes are depicted. Each point of each diagram, with the exception of those points on the symmetry axis (and possibly certain odd points on the boundary), represents a sphere S^2. The symmetry axis (or axes) is depicted by a broken line (or lines), each point of which represents a single point of the spacetime. Past ideal boundary points or future ideal boundary points (normally thought of as distinct from each other—only in the case of anti-de Sitter space, or for the odd point in the extended Reissner-Nordstrom solution, would it be reasonable to "identify" certain past with future boundary points) are depicted by heavy straight lines if they are points at infinity and by jagged lines if they are singular points.

We need a criterion for distinguishing points at infinity from singular points. The simplest is the following (Penrose 1974*a*):

Definition. A TIP is called an ∞-*TIP* if it is of the form $I^-[\gamma]$ for some timelike curve γ of infinite proper length into the future; otherwise it is called a *singular TIP*. The ∞-*TIFs* and singular TIFs are defined correspondingly.

Note that all the TIPs of Minkowski space are ∞-TIPs since they can be generated either by curves of uniform acceleration (for the points of $\mathscr{I}^+$) or timelike straight lines (for i^+), which are all of infinite proper length into the future. It may be remarked that any TIP in any spacetime can always be generated by some curve of *finite* proper length (for example, the future-finite curve $x = 0$, $y = 0$, $z + t = (z - t)^{-1/2} > 0$ in Minkowski space generates the TIP $t < z$) or, indeed, by a curve of zero proper length. However, singular TIPs can be generated *only* by such future-finite curves. So, roughly speaking, *all* observers reach the singular (future) ideal points in finite proper time, whereas the points at infinity can be "reached" by *some* observers only after infinite proper time.

There are certain alternative definitions for distinguishing points at infinity from singular points. Though apparently equivalent to the above in most "normal" circumstances, this need not be the case in every spacetime. One such alternative is the following:

Definition. A TIP (resp. TIF) is called *null-finite* if it is of the form $I^-[\gamma]$ (resp. $I^+[\gamma]$), where γ is a null geodesic of finite affine length; otherwise it is called *null-infinite*.

Each null-finite TIP and TIF may be thought of as defining a sort of singular ideal point. In fact, a null-finite TIP can be an ∞-TIP only if the curvature, in a certain well defined sense, is unbounded along the relevant null geodesic γ. So it would not be unreasonable to think of these TIPs as defining singular points, though perhaps they are best thought of as "singular points at infinity." One

Singularities of spacetime. *Theoretical principles in astrophysics and relativity.* Eds N. R. Lebovitz, W. E. Reid, and P. O. Vandervoort. Chicago University Press, 1978

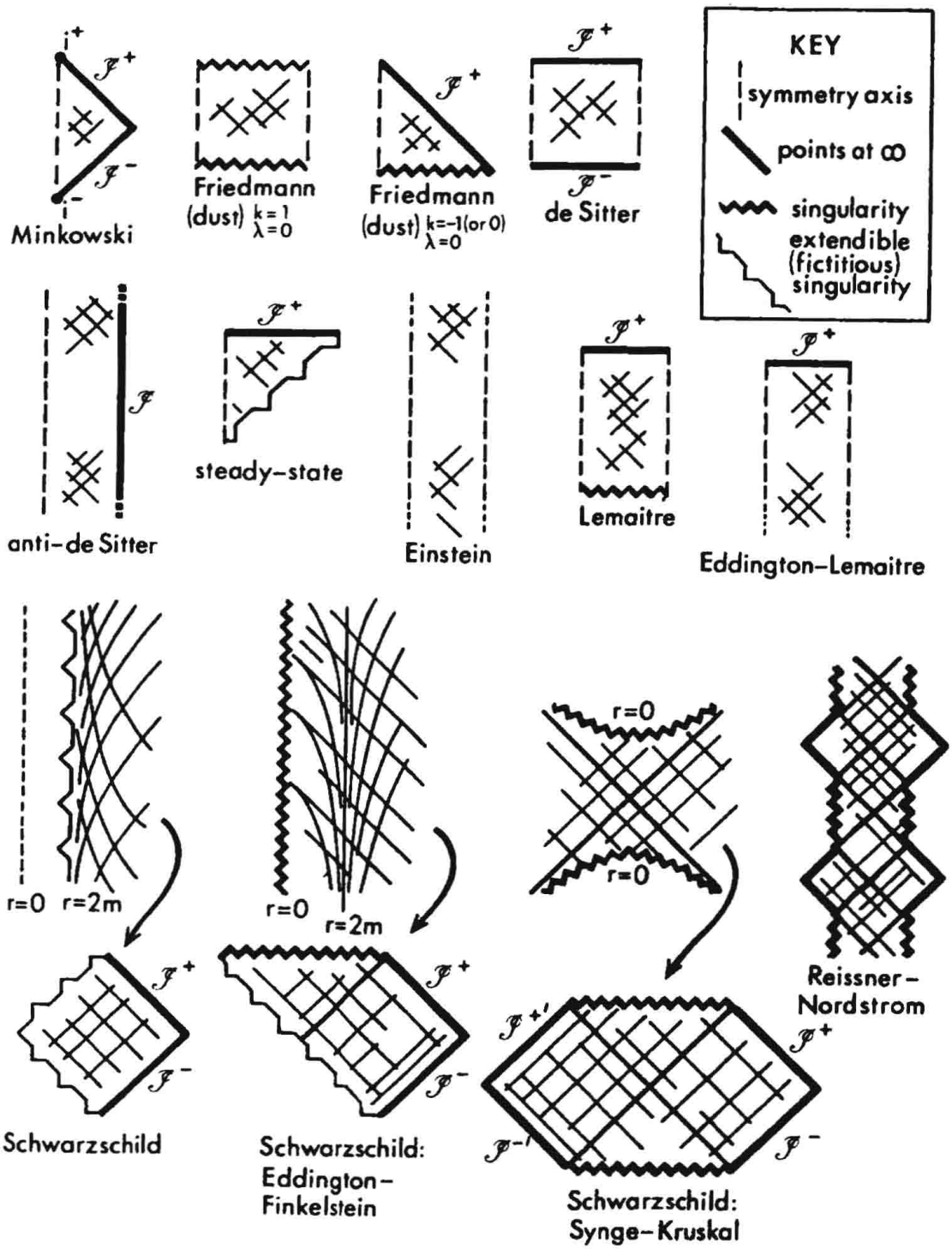


Figure 3. Diagrams illustrating the ideal point structure of various familiar spacetimes.

 Singularities of spacetime. *Theoretical principles in astrophysics and relativity.* Eds N. R. Lebovitz, W. E. Reid, and P. O. Vandervoort. Chicago University Press, 1978

could also envisage certain null-infinite ∞-TIPs representing "singular" points at infinity (because of unbounded curvature along an appropriate generating curve γ). I shall not enter into a general discussion of this matter here, preferring simply to stick to the terminology given in the above definitions.

It is possible to define causal and chronological relations between TIPs and between TIFs in a simple way (Geroch *et al.* 1972). This enables one to discuss the spacelike, null, or timelike nature of the ideal point sets:

Definition. Let P and Q be TIPs, and let F and G be TIFs. Then we say that P *causally precedes* Q if $P \subset Q$ and F *causally precedes* G if $G \subset F$. We say that P *chronologically precedes* Q if $P \subset I^-(x)$ for some $x \in Q$ and F *chronologically precedes* G if $G \subset I^+(x)$ for some $x \in F$.

Causal and chronological relations between TIPs and TIFs can also be defined but I shall not bother with them.

The above definitions can also be applied directly to IPs and IFs generally. In particular, when applied to PIPs and PIFs, it gives causal and chronological relations between points of $\mathfrak{M}$. However, in some spacetimes these are not quite the same as the usual ones (namely, x chronologically precedes y if there is a timelike curve in $\mathfrak{M}$ from x to y; x causally precedes y if there is timelike or null curve in $\mathfrak{M}$ from x to y) but differ from them in that ideal points may now be present in the chronological or causal path connecting the points.

The past boundary of $\mathfrak{M}$ is *acausal* (i.e., completely spacelike) if no two TIFs of $\mathfrak{M}$ are causally related to one another, that is, *no TIF contains another TIF*. It is *achronal* (i.e., completely spacelike or null) if no two TIFs are chronologically related to one another, that is, if *no TIF is contained in a PIF*. (Any PIF $I^+(x)$ is contained in some TIF, namely, $I^+[\gamma|$ where γ extends endlessly into the past from x. Thus if the TIF G is contained in the PIF $I^+(x)$, then $I^+[\gamma]$ chronologically precedes G; that is, we have two TIFs chronologically related.)

Thus, roughly speaking, the system of TIFs would be characterized as *timelike* by the fact that each TIF is contained in a PIF, as *null* by the fact that no TIF is contained in a PIF but each TIF is contained in another TIF, and as *spacelike* by the fact that no TIF is contained in any other IF. The situation for TIPs is similar.

3. Gravitational Collapse

Let us consider the standard picture (Penrose 1969, 1974*a*) of collapse to a black hole and see how the preceding ideas can be applied. First, there is the spherically symmetrical situation where a spherical body collapses inward, its external field being described by the Schwarzschild solution. The body collapses

within its Schwarzschild radius $r = 2m$ and continues right down until the singularity at $r = 0$ is reached. The external field continues to within $r = 2m$, but the original static Schwarzschild coordinates cannot be retained there. Instead, a nonstatic coordinate must be used, such as those of Eddington-Finkelstein for which the metric becomes

$$ds^2 = \left(1 - \frac{2m}{r}\right) dv^2 - 2dvdr - r^2(d\theta^2 + \sin^2\theta d\phi^2)$$

where v is an advanced time parameter.

The situation is depicted in Figure 4, where the representation with null lines

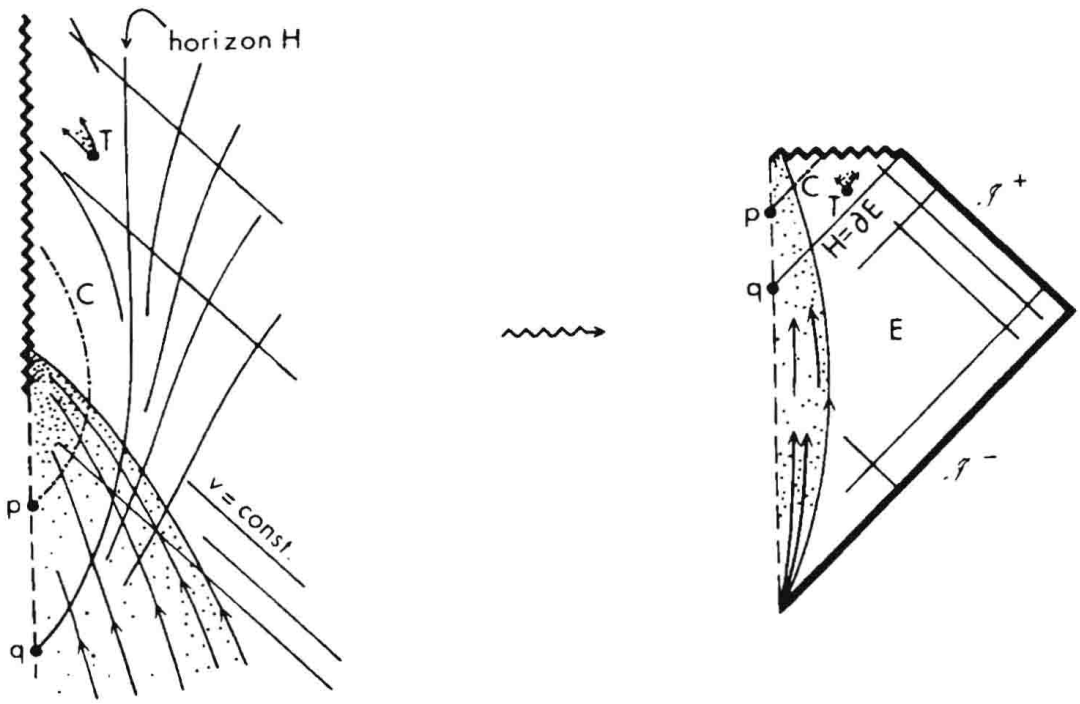


Figure 4. The standard picture of spherically symmetric collapse to a black hole.

at 45° is also given. The TIFs of the spacetime form a system essentially identical with that for Minkowski space. Thus, there is an ideal point i^- (which supplies a past endpoint to any timelike geodesic that is past-endless in $\mathfrak{M}$) corresponding to the TIF $\mathfrak{M}$; and there is a three-dimensional system $\mathscr{I}^-$ of ideal points (supplying past endpoints to the null geodesics past-endless in $\mathfrak{M}$) which is null in the sense described in the last paragraph of § 2. All the TIFs are ∞-TIFs (and null-infinite).

As to the TIPs, we have, similarly to the case of Minkowski space, a null

 Singularities of spacetime. *Theoretical principles in astrophysics and relativity.* Eds N. R. Lebovitz, W. E. Reid, and P. O. Vandervoort. Chicago University Press, 1978

three-dimensional system of ∞-TIPs defining the set $\mathscr{I}^+$ of ideal points of future endpoints of, for example, the various outgoing radial null geodesics. There is also a single ∞-TIP $E = I^-(i^+)$ which is actually the union of all the other ∞-TIPs. The ideal point i^+ is the future endpoint of each timelike geodesic which has infinite proper length into the future. It is also the future endpoint of, for example, the null geodesics which orbit endlessly about the final black hole at $r = 3m$, or of the null geodesic generators of the black hole's

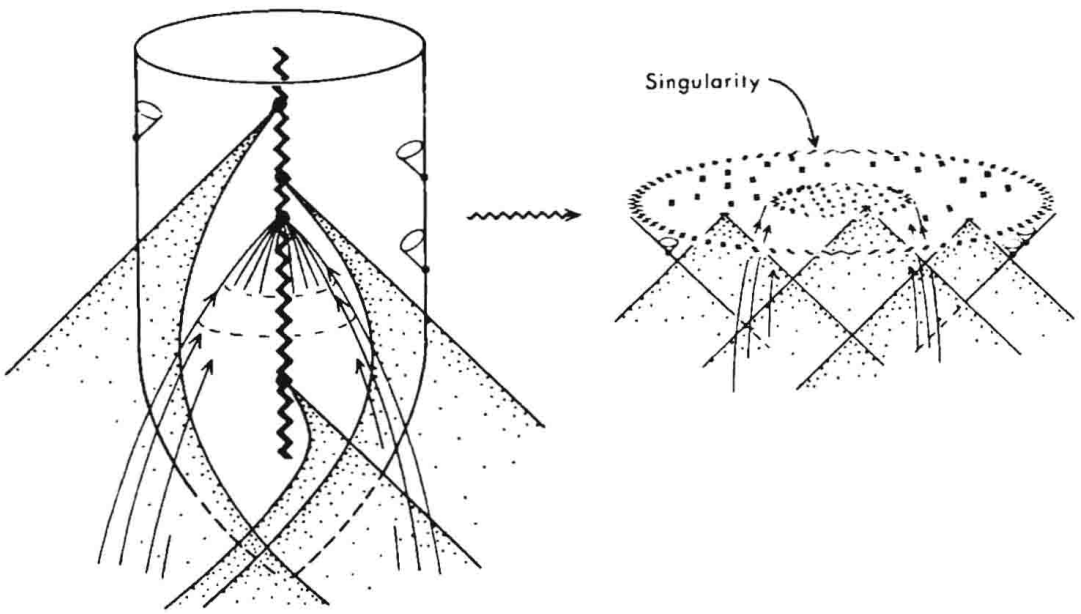


Figure 5. The singular TIPs in spherical collapse define a three-dimensional spacelike set.

absolute event horizon. The ideal point i^+ lies causally to the future of all the points of $\mathscr{I}^+$.

There are also some singular TIPs. These form a spacelike three-dimensional system. This may be ascertained by examining in detail the behavior of the null cones of the Eddington-Finkelstein metric in the neighborhood of $r = 0$ (see Fig. 5). The spacelike nature of the singular part of the future boundary implies that no observer falling into the black hole can actually "see" the singularity before he hits it. In fact any two such observers, unless they share the same future ideal endpoint, must fail to observe the final portions of each other's world-lines. Thus we have the situation of (observer dependent) event horizons (boundaries of TIPs)—a situation familiar in, for example, the de Sitter cosmology. These event horizons do not, of course, coincide with the *absolute event horizon* (Penrose 1969) H, which is the boundary of E:

$$H = \partial E = \partial I^-(i^+) = \partial I^-[\mathscr{I}^+] .$$

In the vacuum region, ∂E consists of the Schwarzschild hypersurface $r = 2m$, while in the matter region it is a null hypersurface—the light cone of some point q at the center of the collapsing body—which fits smoothly on $r = 2m$ at the surface of the body.

This is the standard collapse picture when exact spherical symmetry is assumed. If deviations from sphericity are small, then, since the perturbation analysis of the Schwarzschild solution (Regge and Wheeler 1957; Vishveshwara 1970; Chandrasekhar 1975) indicates stability for spherical black holes, we do not expect any significant deviations from the spherically symmetrical picture—at least in the region from the neighborhood of $r = 2m$ outward. On the other hand, if deviations from sphericity are moderately large, then we cannot rely on perturbation analysis. We can, however, appeal to the singularity theorems to derive the fact that a spacetime singularity, analogous to $r = 0$ in the symmetrical case, will still exist, provided that a suitable criterion of "irreversible collapse" has been met. For such a criterion we recall one of the singularity theorems:

Theorem (Hawking and Penrose 1970). *If a spacetime* $\mathfrak{M}$ *is such that:*

a) the timelike convergence condition[1]

$$R_{ab}t^a t^b \leq 0 \text{ for all timelike } t^a$$

holds at every point of $\mathfrak{M}$,

b) there are no closed timelike curves in $\mathfrak{M}$,

c) the generality condition that[2]

$$l_{[a}R_{b]cd[e}l_{f]}l^c l^d \neq 0$$

holds somewhere along every timelike or null geodesic, its tangent vector being l^a, *is valid in* $\mathfrak{M}$,

d) either (i) $\mathfrak{M}$ *contains a closed spacelike hypersurface* S

or (ii) $\mathfrak{M}$ *contains a trapped surface* T (*a closed spacelike two-surface whose null normals are everywhere converging into the future*)

or (iii) $\mathfrak{M}$ *contains a point* p *such that the divergence of the system* C *of null geodesics through* p *changes sign somewhere to the future of* p;

then $\mathfrak{M}$ *is singular in the sense of containing a future- or past-endless timelike or null geodesic of finite affine length.*

There is another version of this theorem of more relevance to the definitions given here, which can be obtained by suitably modifying the proof of the original version. This is:

[1] If Einstein's equations without cosmological term $R_{ab} - \frac{1}{2}Rg_{ab} = -8\pi T_{ab}$ hold, then this is called the *strong energy condition* (Hawking and Ellis 1973) (or simply the energy condition), namely, $T_{ab}t^a t^b \geq \frac{1}{2}Tt_a t^a$ for all timelike t^a. The Ricci tensor sign is chosen consistent with this.

[2] Square brackets denote skew-symmetrization.

 Singularities of spacetime. *Theoretical principles in astrophysics and relativity.* Eds N. R. Lebovitz, W. E. Reid, and P. O. Vandervoort. Chicago University Press, 1978

Theorem (Penrose 1975). *If a spacetime* $\mathfrak{M}$ *is such that conditions* (*a*), (*b*), (*c*), *and* (*d*) *of the previous theorem hold, then* $\mathfrak{M}$ *is singular in the sense of containing a singular TIP or a singular TIF.*

In these results, condition (i) of (*d*) is relevant only to closed cosmological models, while either (ii) or (iii) can be used as a criterion characterizing irreversible collapse. Referring to Figure 4 for the spherically symmetrical case, we find that a two-surface T lying in the vacuum region and defined by $v = \text{const.}$, $r = \text{const.} < 2m$, is an example of a trapped surface. Its null normals are the pairs of null geodesics through T lying in the various ($\theta = \text{const.}$, $\phi = \text{const.}$)-planes. Since both these systems of null geodesics proceed into the future to decreasing values of r, it follows that their areas of cross-section must be decreasing in future directions (since each sphere $v = \text{const.}$, $r = \text{const.}$, has area $4\pi r^2$), that is, these null normals are converging, in accordance with the definition of a trapped surface. Similarly the future light cone C of the point p indicated in Figure 4 moves in toward decreasing values of r after initially diverging away from p. Thus C satisfies the condition required of it for (iii).

Now, the point about the properties (ii) and (iii) is that they are inequalities of the kind whose validity we would expect to be unaffected if a fairly small, but finite, change is made in the spacetime geometry. Thus we can envisage a collapse which deviates, to a certain degree, from exact spherical symmetry; and we would still expect (ii) or (iii) to hold. The geometry near $r = 0$, where the singularity occurs in the symmetrical case, might well be drastically changed. But our theorem tells us that while the detailed structure of the singularity may now be quite different, a spacetime singularity of some sort must still be present. This, of course, is supposing that the remaining requirements (*a*), (*b*), and (*c*) still hold. In fact, they are reasonable requirements for a classical spacetime described according to general relativity, where (*c*) would be expected to hold in all but a very few special cases, of "measure zero" among the totality of possible models. [Curiously, though, (*c*) tends to fail for the known exact solutions. In particular, it fails for the spherically symmetrical collapse just considered!]

The use of (iii) as a criterion of irreversible collapse is perhaps the easiest to visualize. We imagine matter (perhaps a cloud of gas or stellar material, perhaps a collection of stars, or even of galaxies) falling inward in the general direction of some central point p. A flash of (idealized) light is emitted from p, which travels outward describing the spacetime cone C. The infalling matter crosses C and in doing so will cause a certain amount of focusing of the null rays in C. (Of course, this focusing effect on light rays passing near or "through" a material body is, in the case of the Sun, one of the *observed* effects of general relativity.) If sufficient matter crosses C in a sufficiently small region (a condition easily met with quite small densities, provided the total mass is large enough),

then the divergence of the null rays in C will change to a convergence in each direction and condition (iii) will be satisfied.

Once this happens, we know from the theorem that we have a singularity. But we do not yet know whether we have an absolute event horizon enclosing the singular region and preventing, as in the spherically symmetrical case first considered, any information escaping from the neighborhood of the singularity to external infinity. If a singularity is visible[3] from infinity, it is called a *naked singularity*. If there is a region near the singularity which is not visible from infinity, then we have a (nonvacuous) *absolute event horizon*, namely, the boundary $H = \partial E$ of the set E of events from which timelike curves to infinity exist. There is also the possibility that there might be both a naked singularity *and* an absolute event horizon. But if there is only the horizon and no naked singularity, this is the situation of a *black hole* (or holes). Thus, we may conclude, roughly speaking, that if irreversible collapse takes place (and general situations can be set up so that it does), then it must result in a black hole (or holes) *unless* a naked singularity occurs—an alternative possibility which would normally be viewed with considerably more alarm than that of a black hole! For whatever unknown physics actually takes place at a spacetime singularity, its effects would be relevant *observationally* if and only if the singularity is a visible one.

4. Cosmic Censorship

The hypothesis that a physically realistic collapse will not result in naked singularities is called *cosmic censorship* (Penrose 1969). It is not known at present whether some form of cosmic censorship principle is actually a consequence of general relativity. In fact, we may regard this as possibly the most important unsolved problem of classical general relativity theory. The question is not simply whether solutions of Einstein's equations exist which possess naked singularities, or even whether solutions exist which represent the collapse of a perfectly regular initial configuration to a naked singularity. Various such solutions are already known, but they represent special situations that we have no reason to believe would arise in a realistic collapse. It is, rather, whether such collapse configurations lead *stably* to the production of naked singularities, the equations of state for the collapsing matter also not being, in some appropriate sense, too special or physically unrealistic.

Indeed, it may well be that the framing of an appropriate mathematical statement of the cosmic censorship principle could represent, in itself, a major step toward a solution of the problem. The examination of particular exact models would seem to be of rather limited value. What is required is an under-

[3] The term "visible" is here being used somewhat loosely, but meaning, roughly, that a timelike or null curve exists from the singularity to external infinity.

standing of the generic case. It would not, it seems, even be greatly helpful to have a full appreciation of the generic static or stationary solution of Einstein's equations. We know from the Israel-Carter-Hawking-Robinson theorem (Israel 1967; Carter 1971, Hawking 1972; Robinson 1975) (essentially) that the stationary asymptotically flat vacuum spacetimes which are not described by Kerr metrics (with $a \leq m$) cannot be black holes; so presumably these all have naked singularities. From this point of view it is not surprising that the various (non-Kerr) stationary asymptotically flat vacuum solutions which have been proposed at various times (Weyl - Levi Civita, cf. Curzon 1924; Tomimatsu-Sato 1972; Witten 1974; etc.) all possess naked singularities. They tell us little, if anything, about the validity of cosmic censorship. In those circumstances in which a collapse does result in a black hole, the process of settling down to a stationary state leads to all multipole moments of quadrupole or higher character becoming adjusted to those values given by the Kerr field (so it is to be expected) with the emission of gravitational radiation. So *if* cosmic censorship is true, these non-Kerr stationary solutions have no direct relevance.

As regards the nonstationary asymptotically flat collapse models with naked singularities (e.g., Yodzis *et al.* 1973; Steinmüller *et al.* 1975; Szekeres 1975; cf. also Grischuk 1967), two main queries refer to the equations of state and to the generality of the solutions. The use of a pressureless fluid (dust) in some of these models raises some questions since we expect caustics, and hence infinite densities, in the fluid world-lines. While it might be that a very soft equation of state (e.g., that of Hagedorn 1968) has relevance when the situation is extreme enough, the question must be raised as to why the situation should become that extreme in the first place unless the initial setup involved a considerable amount of delicate (and intrinsically improbable) careful alignment. This brings us back to the question of generality. It would seem that this is really the key issue. Exact models might technically possess naked singularities which could be argued to be of a "harmless" kind (like those arising in other fields of physics where delta-functions are employed as a mathematical convenience, not being supposed to imply that the physical quantities that they describe are actually infinite). The issue at hand is whether such singularities would persist when the initial conditions—*and* equations of state—are replaced by something appropriately generic. It is not yet clear how this could be satisfactorily formulated, particularly as regards the equations of state. However, we might hope to express the general question in the form: "Is the production of naked singularities in a collapse a stable property?" If it never is, then we could say that the cosmic censorship hypothesis is valid.

There is a possible line of attack for obtaining a contradiction with the cosmic censorship hypothesis which depends on the fact that one of its implications, appropriately stated, is the *principle of area increase* for the absolute event horizon. The first published precise version of this principle was given by Hawk-

ing (1972). Some alternative versions (which are rather more general) will be stated shortly. By use of the area principle we can infer that certain inequalities must hold which relate the geometry of the initial setup to the surface area of the final black hole. This area, in turn, relates to the mass of the final configuration; and since this mass cannot exceed the initial mass, we arrive at a consistency condition on the initial geometry. In appropriate circumstances (e.g., collapsing shells of radiation [Penrose 1973], time-symmetric initial state for black holes [Gibbons 1972], colliding extreme relativistic particles or black holes [Penrose 1976]) this can lead to precise mathematical inequalities whose validity is a consequence of cosmic censorship. In no case has a situation been yet found in which these inequalities are violated, whereas in some fairly general circumstances proofs of the inequalities have actually been obtained (Gibbons 1972; Penrose 1976; Jang 1976). This may be regarded as providing some indirect support for cosmic censorship.

However, it could also be argued that the cases which have so far been handled have simply not been concerned with situations which are aspherical enough for cosmic censorship to be violated. As the stability analysis of black holes (Teukolsky 1973; Press and Teukolsky 1973; Hartle and Wilkins 1974; Chandrasekhar and Detweiler 1976; Stewart 1975) already appears to show (and as do arguments involving test fields on black hole backgrounds [Penrose 1973]), one might in any case expect that any violation of cosmic censorship would have to involve a collapse configuration rather far from the original spherically symmetrical exact model. One possibility, for example, is that sufficient angular momentum would be required, corresponding to a value given by a Kerr metric with $a > m$, that is, with angular momentum (in gravitational units) larger than the mass squared. If such a configuration were able, without loss of angular momentum, to collapse down to a Kerr ($a > m$) state (with central ring singularity spanned, say, by a disk of matter), then we should have a naked singularity. This is an entirely speculative possibility. But it serves to illustrate that catastrophic gravitational collapse to a naked singularity cannot, at the moment, be ruled out. And there seems no reason why even conditions (*d*)(ii) and (*d*)(iii) of the singularity theorems of § 3 need ever be satisfied for such a collapse.

As mentioned earlier, even a clear mathematical statement of the cosmic censorship hypothesis is at present lacking. On the other hand, if we ask merely for a criterion which can be applied to a *given* spacetime that it be free of naked singularities, then a number of different definitions can be given. Three of these are Hawking's conditions (Hawking 1972; Hawking and Ellis 1973) of *future asymptotic predictability*, *strong future asymptotic predictability*, and *regular predictability* (in order of increasing restrictiveness, other properties than just absence of naked singularities being also incorporated). Hawking shows that in a regular predictable spacetime the principle of area increase must hold. The

 Singularities of spacetime. *Theoretical principles in astrophysics and relativity.* Eds N. R. Lebovitz, W. E. Reid, and P. O. Vandervoort. Chicago University Press, 1978

conditions are intended to apply only to spacetimes which are asymptotically flat, and they become a little complicated to state precisely. I shall not bother to give them in detail here (cf. Hawking and Ellis 1973), but suggest some other criteria which are somewhat simpler to state and of considerably wider applicability.

The first point to make is that if a hypothesis for excluding naked singularities is to be adopted, this hypothesis should not exclude singularities of the big-bang type—for otherwise one would presumably be ruling out the actual universe![4] Thus, although we expect the big-bang singularity to be "visible," it should not be classed as a proper naked singularity. We may make a distinction by requiring that a true naked singularity q be "naked" also in the time-reversed sense (Penrose 1974*a*, *b*). Thus, not only shall we require some point r lying to the future of the singularity q, but also some other point p lying to the past of q. Whether these past and future relations are to be taken in the causal or chronological sense makes no essential difference. (The exact locations of p and r can be redefined if necessary to accommodate this.) Since there are no points to the past of the big bang (according to normal ideas), it is *not* a naked singularity. The singularity q may be described as a singular TIP or else a singular TIF. In the former case (taking q to precede r causally) we have this TIP contained in the PIP $I^-(r)$. The TIP (by virtue of the fact that it *is* a TIP) will also have points lying to its (chronological) past. So the condition that this TIP represent a naked singularity is that it be a *singular TIP contained in a PIP*. Similarly, the condition that a TIF represent a naked singularity is that it be a *singular TIF contained in a PIF*.

Two possible statements which state absence of naked singularities for the spacetime $\mathfrak{M}$ are, therefore,

CCI: no PIP contains a singular TIP

and

CC2: no PIF contains a singular TIF.

We might also choose to exclude "naked points at infinity" to obtain the stronger statement (Penrose 1974*a*)

CC3: no PIP contains a TIP,

which is actually *equivalent* to its time-reverse

CC3: no PIF contains a TIF

because each can be shown to be equivalent to the time-symmetric statement

CC3: $\mathfrak{M}$ is globally hyperbolic.

[4] Condition (*d*)(iii) of the singularity theorems can be used in time-reversed form, where it states that the past null cone of some point p is refocused, i.e., that small objects of fixed intrinsic size reach a minimum angular diameter, as viewed from p, when at a certain distance from p. This is actually a condition whose observational feasibility is seriously considered by cosmologists. If satisfied, it implies that the big-bang singularity will persist even after generic perturbation.

A spacetime $\mathfrak{M}$ is said to be globally hyperbolic (Hawking and Ellis 1973; Penrose 1972; Geroch 1970) if it is strongly causal and if $I^+(x) \cap I^-(y)$ has compact closure for each x, $y \in \mathfrak{M}$. By a theorem of Geroch (1970), this is equivalent again to

CC3: $\mathfrak{M}$ admits a global Cauchy hypersurface.

Using the definitions at the end of § 2, we can reinterpret CC1 as the statement that the future singular boundary of $\mathfrak{M}$ is nowhere timelike, and CC2 as the statement that the past singular boundary is nowhere timelike. Thus, we may envisage cosmic censorship as stating something like:

"timelike singularities are unstable,"

so that in a generic collapse we would not expect timelike singularities to occur at all.[5]

In this form, the cosmic censorship hypothesis is rather stronger and more "local" than versions which depend on observations of singularities from infinity. But I feel that it offers more scope for possibilities of finding a general proof. One would expect, after all, that the formation of a singularity should be governed by reasonably local matters and not by how it is related to infinity.

The singularity formed at the center of the spherically symmetrical collapse described in § 3 is spacelike, as we have seen, so CC1, CC2, and, indeed, CC3 are all satisfied. When spherical symmetry is not assumed, the matter needs more careful consideration, however. In fact, for the exact Kerr metric with $a < m$ we have a spacetime rather similar to that of Reissner-Nordstrom, as depicted in Figure 3, and CC1 and CC2 are both violated. Thus, a suitable observer who falls into such a black hole can see the singularity; to *him* the singularity is actually naked (see Fig. 6). But there are reasons for believing that the spacetime and its singularity structure inside the event horizon are unstable (Simpson and Penrose 1973) (despite the probable stability of the spacetime from the absolute event horizon outward). It seems likely that a generically perturbed Kerr or Reissner-Nordstrom solution develops a spacelike (or null) singularity in the neighborhood of its "inner" (or Cauchy) horizon (Hawking and Ellis 1973; Penrose 1972) (see Fig. 6), so that CC3 becomes valid. (If this is the case, then the possibility, sometimes envisaged, of escaping to a "new universe" by entering a rotating black hole must be ruled out.)

If a cosmic censorship principle along this sort of line could be established, then even that would not be a completely satisfactory result. For suppose a gravitational collapse takes place which results in a singularity whose effects are visible from infinity, but where, unlike the comparatively mild type of naked singularity considered earlier, this singularity emits a "pulse" of gravitational

[5] The naked ring singularity of the $a > m$ Kerr metric may repay examination. Since rather few null rays escape the singularity, there is the chance that it might not decay as rapidly as "normal" timelike singularities.

 Singularities of spacetime. *Theoretical principles in astrophysics and relativity.* Eds N. R. Lebovitz, W. E. Reid, and P. O. Vandervoort. Chicago University Press, 1978

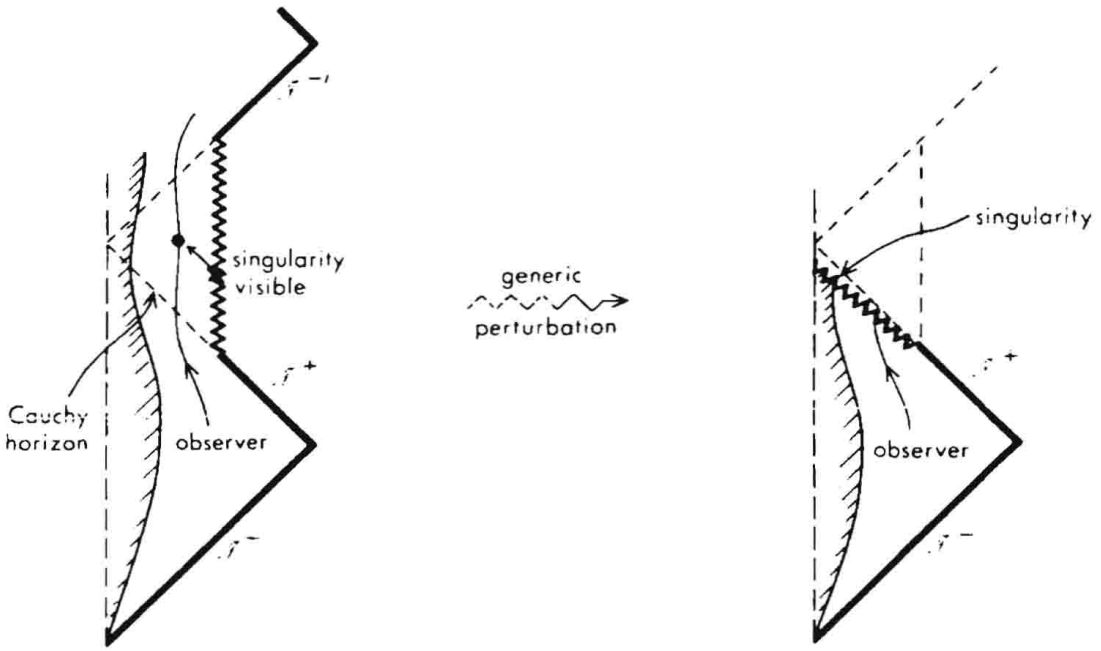


Figure 6. An observer crossing the Cauchy horizon in charged spherical collapse sees an apparently naked singularity. But this situation seems unstable against generic perturbation.

radiation of infinite amplitude—a null shell of singular curvature which destroys the universe as it expands! This should be the "naked singularity to end all naked singularities"—but nevertheless CC3 could be satisfied for the spacetime (see Fig. 7). A hypothesis to exclude such a possibility would be:

CC4: no ∞-TIP contains a singular TIP;

and we might also postulate its time-reverse

CC5: no ∞-TIF contains a singular TIF.

A more local version of these would be to exclude *null* singular boundary points by requiring (in addition)

CC6: no singular TIP contains another singular TIP

and

CC7: no singular TIF contains another singular TIF.

As a test of the utility of such postulates, we may ask whether they imply a principle of area increase for the absolute event horizon. In fact there is a variety of such results. I shall only state two of these here. First, a general definition of absolute event horizon H is required:

$$H = \partial E ,$$

where E is the "external region." We have some choice for the precise definition

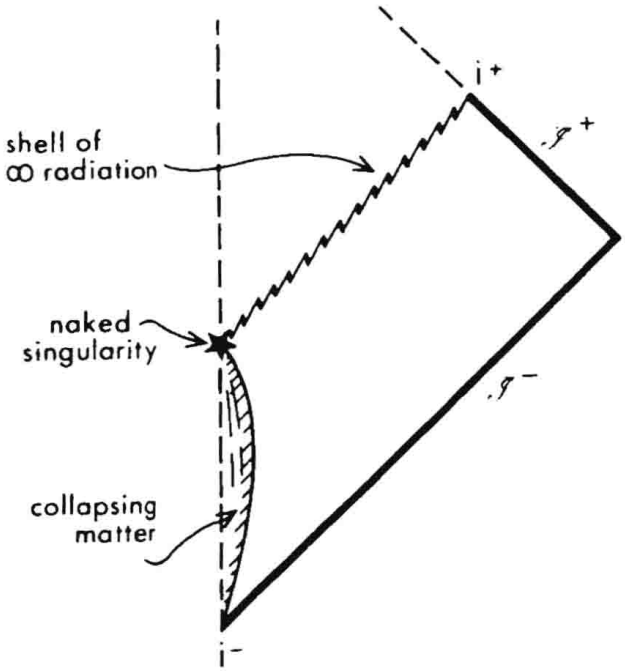


Figure 7. A hypothetical situation in which a "naked" singularity is formed emitting a shell of radiation of infinite intensity--which destroys the universe!

of E. The simplest would seem to be

E = set of all past endpoints of timelike curves of infinite length into the future

= union of set of all ∞-TIPs.

One might prefer, on the other hand, simply to single out some suitable subset of the ∞-TIPs, or perhaps of the null-infinite TIPs, and designate them as "external", calling them *E-TIPs*,[6] their union being, by definition, the set E. For the present purposes it makes no difference, although the characterization "absolute" for the event horizon would, in the latter case, need justification.

If S_1 and S_2 are two spacelike two-surfaces on H (possibly disconnected, but normally cross-sections of H), we can say that *S_2 lies to the future of S_1 on H* if S_1 is in the causal past of S_2. Since H is an achronal set (being the boundary of a future-set), this means that each point of S_1 is the past end-point of a null geodesic on H meeting S_2. The following result has been established (Penrose 1974*a*, 1975):

Theorem. *If $\mathfrak{M}$ satisfies CC3 and the timelike convergence condition (a), and if S_2 lies to the future of S_1 on H, then*

$$area(S_2) \geq area(S_1) .$$

[6] For example, those corresponding to the points of $\mathscr{I}^-$, for a weakly asymptotically simple spacetime. Then (Penrose 1969) $E = I^-[\mathscr{I}^+]$.

 Singularities of spacetime. *Theoretical principles in astrophysics and relativity.* Eds N. R. Lebovitz, W. E. Reid, and P. O. Vandervoort. Chicago University Press, 1978

Another version of the area principle[7] which is considerably easier to establish uses a modified form of CC4, namely,

CC8: no E-TIP contains a null-finite TIP.

Then one can show:

Theorem. *If $\mathfrak{M}$ satisfies CC8 and the null convergence condition,[8] and if S_2 lies to the future of S_1 on H, then*

$$area(S_2) \geq area(S_1)$$

5. Quantum Effects and Time-Asymmetry

The discussion to this point has been entirely classical. In the normal way, one does not expect quantum-mechanical effects to have significance in gravitational theory owing to the enormous disparity in the scales involved—gravitation being normally relevant only for masses much larger than 10^{-5} g and quantum mechanics for masses much smaller than 10^{-5} g. However, one expects the presence of large curvatures near singularities—and quantum effects relating to gravity are likely to be important there. This could well be the case even in situations where the spacetime can still be treated as a classical background, the matter fields being described quantum-mechanically. For then there can be local violation of energy conditions, in addition to particle production effects. One might expect that when averaged out on a larger scale these would not *normally* lead to violation of the timelike convergence condition. But it is in any case apparent that the various classical results like the singularity theorems and area-increase principle can be trusted to have detailed validity only to the extent that such small-scale violations can be ignored.

These remarks have relevance to the question of the behavior of mini-black holes, which Hawking suggested several years ago (Hawking 1971) could be produced in a sufficiently chaotic initial big bang. They might have masses down to 10^{15} g (radius 10^{-13} cm) or even down to 10^{-5} g (radius 10^{-33} cm). More recently, Hawking has shown (Isham *et al.* 1975; Hawking 1975) that a remarkable phenomenon occurs, of significance for such mini-black holes, namely, that they would emit blackbody radiation at a temperature inversely proportional to the black-hole mass. This effect would also be present for large black holes, but utterly insignificant for ordinary astrophysics, the temperature associated with a solar-mass black hole being less than 10^{-7} K.

But a hole initially small enough that its temperature exceeds the tempera-

[7] Yet a different version has just recently been given by Gibbons and Hawking (1975).

[8] This is a weakened version of the timelike convergence condition (*a*) for which the timelike vector t^a is replaced by a null vector (Hawking and Ellis 1973). It has the advantage that Einstein's equations can be used with cosmological term, and it then becomes the weak energy condition—which is a consequence of the positive-definiteness of local energy.

ture of the background radiation in which it is immersed would emit more radiation than it absorbs. As it does so, it loses energy and therefore mass. Thus, its surface area likewise decreases, in violation of the area-increase principle. This can be attributed to a violation of the weak energy condition, in accordance with the fact that this is a situation where the matter fields must be treated quantum-mechanically.

As the mass goes down, the temperature goes up, and the effect is accelerated. For a mass smaller than about 10^{17} g, electrons and positrons would be emitted. If the mass were as small as about 10^{14} g, the temperature would reach 10^{12} K, which, according to some strong interaction models such as that of Hagedorn (1968), should be a maximum temperature. Instead of the temperature rising further, more and more species of elementary particles would be produced, and the black hole would presumably disappear in an almost instantaneous explosion (10^{35} ergs in about 10^{-23} s). Alternatively, if there were no such maximum temperature, the black hole might continue to reduce in size and mass until it reached a size of 10^{-33} cm—the Planck scale, at which quantum fluctuations in the metric would be large enough to preclude any classically meaningful spacetime geometry. The lifetime for a hole of mass 10^{9} g would be about a tenth of a second, and 10^{30} ergs would be emitted. For 10^{14} g the lifetime would be 10^{7} years and for 10^{15} g, a Hubble time (being inversely proportional to the cube of the mass).

For those, such as myself, who would like to believe that radii of spacetime curvature smaller than the strong-interaction scale of about 10^{-13} cm (or so) do not make consistent physical sense, there is a certain appeal in the Hagedorn model. A black hole of 10^{-14} cm would evaporate essentially instantaneously! At this final stage, the normal spacetime notions would be violated and we have the situation of a naked singularity.[9] Thus, with the Hagedorn strong-interaction model, this singularity already arises at the regime where radii of spacetime curvature are of the general order of the strong-interaction length.

In Figure 8, spacetime diagrams (analogous to those of Fig. 4) are given for the entire process of formation and final evaporation of the black hole. It is being assumed that the final naked singularity is ephemeral, no residual singularity of any kind being left in the spacetime which persists with time. There is just one singular TIF, and it violates CC2. Of the singular TIPs, apparently an S^2's worth violates CC1. Nevertheless, this naked singularity[10] cannot really be considered to be timelike. We note that these TIPs also violate CC4. However,

[9] This is not considered to be in violation of the cosmic censorship hypothesis, which is concerned with *classical* general relativity only.

[10] Although represented by a single point in Fig. 8, this naked singularity appears to have a different structure as a TIF than as a set of TIPs (cf. Geroch *et al.* 1972).

 Singularities of spacetime. *Theoretical principles in astrophysics and relativity.* Eds N. R. Lebovitz, W. E. Reid, and P. O. Vandervoort. Chicago University Press, 1978

CC6 and CC7 are both satisfied for the whole spacetime, showing that by themselves they are insufficient to characterize absence of naked singularities.

Whether or not mini-black holes would be expected to exist in the actual universe depends on how chaotic the initial big bang was. With a soft equation of state (effectively zero pressure at high densities) such as that implied by the Hagedorn model, mini-black holes would be produced with a comparatively uniform (i.e., unchaotic) big bang. If the Hawking explosions are not observed (and some strong observational limits on this are already at hand [Carr 1975]), then this may be construed as implying a high degree of small-scale uniformity for the big bang. In fact, I shall attempt to present a completely different theoretical reason for believing in a uniform big-bang singularity. The main objection to this point of view, on the other hand, comes from the problem of galaxy formation. No plausible mechanisms have yet been put forward which do not trace back the origins of galaxies to some initial irregularities of the big bang. My own view on this is that we should wait and see if some alternative proposal for galaxy formation may not eventually be forthcoming.

The main interest in the Hawking evaporation process is, in my view, a purely theoretical one. The idea had already been put forward by Bekenstein

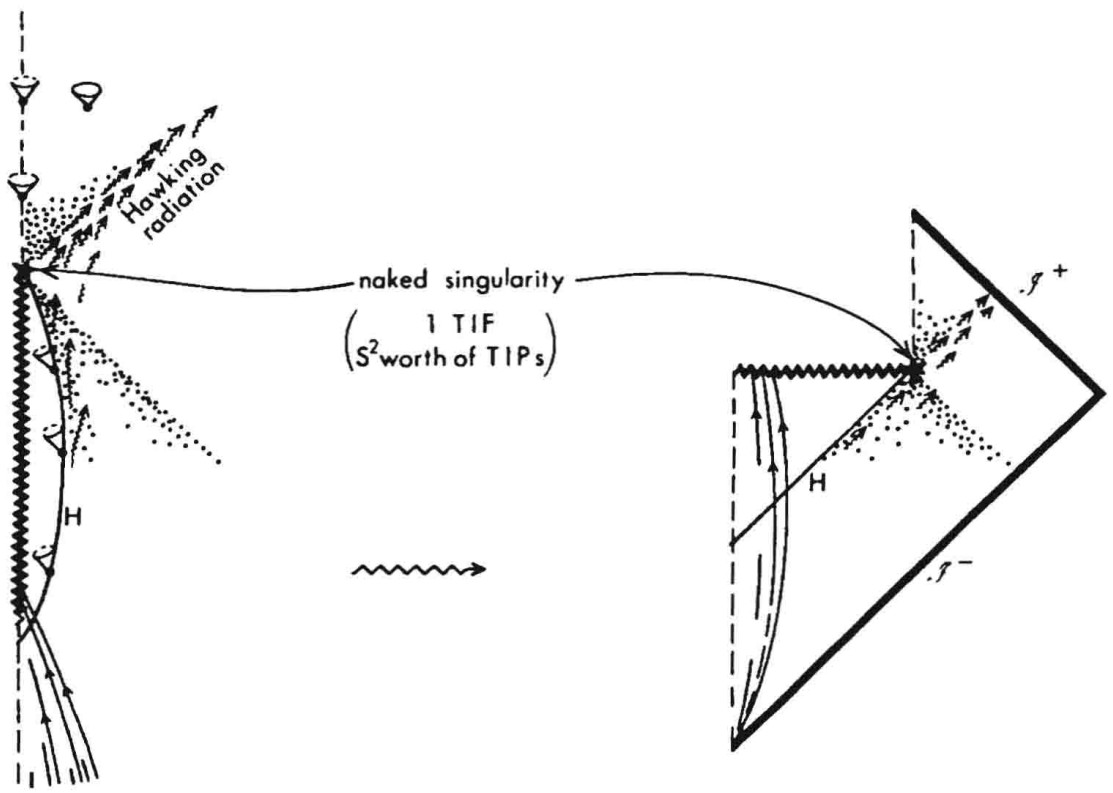


Figure 8. Spherical collapse to a black hole which subsequently evaporates via the Hawking process to a naked singularity. The ideal point structure is illustrated.

(1973, 1974) that an entropy

$$S_{\mathrm{bh}} = \frac{1}{4}\frac{k}{\hbar}A$$

(in gravitational units, k being Boltzmann's constant, the precise value of the coefficient $\frac{1}{4}$ having been supplied later by Hawking [Isham *et al.* 1975; Hawking 1975]) should be attributed to any stationary black hole of surface area A. This he makes very plausible by means of thought experiments, where one attempts to violate the second law of thermodynamics by lowering high-entropy matter or radiation into the hole. We note the presence of $\hbar$ in this formula, which is essential for dimensional reasons. Thus, such an identification of a black-hole entropy with the area of the horizon makes physical sense only by virtue of the fact that it is *quantum* physics which is involved.

This identification of entropy leads directly (Bardeen *et al.* 1973) to the association of a *temperature* with the quantity (inversely proportional to the black hole's mass in the case of a nonrotating hole) known as the black hole's *surface gravity* (Hawking and Ellis 1973). It is this very temperature which features in the Hawking evaporation process. In fact, the Hawking process is *necessary* for the Bekenstein entropy expression to be thermodynamically consistent. But when these two features are incorporated together, there is then a remarkable elegance and coherence in the whole scheme, linking, as it does, apparently disparate elements from general relativity, thermodynamics, and quantum field theory.

However, there are also some unusual features. In the first place, the black hole has a negative specific heat. This can lead to unstable situations such as that already discussed, where a black hole can get out of equilibrium with its surroundings, the final result of this runaway situation being (in the case considered) a naked singularity. Second, there is a time-asymmetry present in the whole setup. This is manifest particularly in the geometry of Figure 8. Even if we delete the entire portion of the spacetime inside the horizon, we do not get a time-symmetrical picture. Not least, there is the striking fact that an apparently completely objective feature of the spacetime geometry has been put equal to a thermodynamic quantity--which one might have thought depended upon some subjective notion of coarse-graining.

However, despite these features, Hawking has been led to consider the hypothetical state of equilibrium (maximum entropy) of a system of fixed mass, constrained in a box with perfectly reflecting walls. When the ratio of mass to box volume lies within an appropriate range, he finds this state of equilibrium to consist of a black hole in equilibrium with blackbody radiation. He then argues that general thermodynamic considerations should imply that this configuration is time-symmetric, so that a black hole should be physically indistinguishable from a white hole (time-reverse of a black hole). Though the normal accretion process whereby a black hole acquires mass is a purely classical one, and the

 Singularities of spacetime. *Theoretical principles in astrophysics and relativity.* Eds N. R. Lebovitz, W. E. Reid, and P. O. Vandervoort. Chicago University Press, 1978

evaporation process whereby it loses mass is quantum-mechanical, Hawking (1975) argues that they are physically time-reverses of one another!

Striking as these ideas undoubtedly are, it seems to me that they run into severe difficulties when examined in detail. For example, it appears that large fluctuations away from equilibrium do not behave time-symmetrically (Penrose 1976). More alarming is the fact, already mentioned, that the spacetime geometry of Figure 8 is not time-symmetrical, so we are led to the view, if we accept Hawking's arguments in full, that this geometry is not objectively meaningful, even at the level of macroscopic dimensions (say of a solar-mass size black hole a few kilometers across).

My own view is that this remarkable link between black holes and thermodynamics which is supplied by the Bekenstein-Hawking ideas is really trying to tell us something different which may eventually help to solve two of the long standing problems of physics. One of these concerns the objectivity of quantum phenomena and the vexed question of the "reduction of the wave function" when an "observation" is made. I do not propose to enter into a discussion of this here, but merely point out that the question of the objectivity (i.e., observer independence) of spacetime geometry has already been mentioned above in two somewhat different contexts. The other question has to do with the statistical time-asymmetry of physics.

I take the view that, at least in normal physical circumstances and where spacetime curvature radii are much larger than 10^{-13} cm, the geometry of spacetime can be treated objectively. Then black holes and white holes (of mass greater than 10^{15} g) are different. Furthermore, it seems to me reasonable that only the black holes should exist in our universe. Since each represents a valid solution of the Einstein equations in its nonsingular part, the reason for ruling out the white hole would be that its singularity represents an unacceptable boundary condition. The view would be that it is a "high entropy" singularity acceptable only as a final boundary condition and not as an initial one. Singularities of even higher entropy could arise as the result of an asymmetric collapse to a black hole, since in this case even more information is lost in the hole than when spherical symmetry is assumed. The big bang, on the other hand, must be a low-entropy singularity. The initial *matter* distribution is, after all, normally assumed to be in thermodynamic equilibrium. The reason that the universe does not start off in a state of maximum entropy is presumably that the gravitational degrees of freedom are not used (not "thermalized") initially. With gravitation, it is irregular states that have high entropy and the smooth ones that have low entropy (in contradistinction with the normal situation). We can envisage that some suitable integrated measure of the size of the Weyl (conformal) curvature might give a measure of the entropy in the geometry. This could have a high value for the final collapse singularities—where the Weyl curvature is expected not only to diverge to infinity but also to dominate over

the remaining Ricci (matter) part (Belinskii *et al.* 1970). The initial big bang, on the other hand (Zel'dovich 1973), seems to have been very uniform. This corresponds to the vanishing of Weyl curvature for the big bang, which is the case for the Friedmann-Robertson-Walker models, and which more or less serves to characterize their initial singularities.

Thus, one may make the hypothesis that initial singularities (singular TIFs) are for some reason[11] constrained to have zero Weyl curvature, whereas final singularities (singular TIPs) are not so constrained. The past and future boundary conditions for the universe must indeed be very different in structure from one another in order that the observed gross statistical time-asymmetry of our world can be attributed a cause.

References

Bardeen, J. M., Carter, B., and Hawking, S. W. 1973, *Comm. Math. Phys.*, **31**, 170.
Bekenstein, J. C. 1973, *Phys. Rev. D*, **7**, 2333.
———. 1974, *Phys. Rev. D*, **9**, 3292.
Belinskii, V. A., Khalatnikov, I. M., and Lifshitz, E. M. 1970, *Adv. in Phys.*, **19**, 525.
Bosshard, B. 1975, On the *b*-Boundary of the Closed Friedmann Model, preprint.
Carr, B. J. 1975, Caltech Orange Aid preprint No. 415.
Carter, B. 1971, *Phys. Rev. Letters*, **26**, 331.
Chandrasekhar, S. 1975, *Proc. Roy. Soc. London*, **A343**, 289.
Chandrasekhar, S., and Detweiler, S. 1976, *Proc. Roy. Soc. London*, **A345**, 145.
Curzon, H. 1924, *Proc. London Math. Soc.*, Vol. **23**.
Geroch, R. P. 1968, *J. Math. Phys.*, **9**, 450.
———. 1970, *J. Math. Phys.*, **11**, 437.
Geroch, R., Kronheimer, E. H., and Penrose, R. 1972, *Proc. Roy. Soc. London*, **A327**, 545.
Gibbons, G. W. 1972, *Comm. Math. Phys.*, **27**, 87.
Gibbons, G. W., and Hawking, S. W. 1975, preprint.
Grischuk, L. P. 1967, *Soviet Phys.-JETP*, **24**, 320.
Hagedorn, R. 1968, *Nuovo Cimento*, **56A**, 1027.
Hartle, J. B., and Wilkins, D. C. 1974, *Comm. Math. Phys.*, **38**, 47.
Hawking, S. W. 1971, *M.N.R.A.S.*, **152**, 75.

[11] Since the balance of what is important in local physics is presumably greatly different near a singularity from what it is elsewhere, there is the possibility that the interaction responsible for the time-symmetry of K^0 decay may be playing a role (Penrose 1969, 1976). I should emphasize, moreover, that no explanation can be found merely in particle creation effects (Misner 1969; Novikov and Zel'dovich 1973; Lukash and Starobinskii 1974) (say) which depend on a *time-symmetric* local physics, since all arguments would apply equally to the future singularities. Furthermore, it seems that *TCP* violation is required in addition to *T* violation.

———. 1972, *Comm. Math. Phys.*, **25**, 152.
———. 1975, Fundamental Breakdown of Physics (preprint).
Hawking, S. W., and Ellis, G. F. R. 1973, *The Large Scale Structure of Space-time* (Cambridge: Cambridge University Press).
Hawking, S. W., and Penrose, R. 1970, *Proc. Roy. Soc. London*, **A314**, 529.
Isham, C. J., Penrose, R., and Sciama, D. W. 1975, ed., *Quantum Gravity* (Oxford: Oxford University Press).
Israel, W. 1967, *Phys. Rev.*, **164**, 1776.
Jang, P. S. 1976, private communication to A. Ashtekar.
Lukash, V. N., and Starobinskii, A. A. 1974, *Soviet Phys.—JETP*, **39**, 742.
Misner, C. W. 1969, *Phys. Rev.*, **186**, 1319, 1328.
Novikov, I. D., and Zel'dovich, Ya. B. 1973, *Ann. Rev. Astr. Ap.*, **11**, 387.
Penrose, R. 1969, *Rev. Nuovo Cimento*, Ser. 1. Num. Spec., **1**, 252.
———. 1972, *Techniques of Differential Topology in Relativity* (Philadelphia: SIAM).
———. 1973, *Ann. N.Y. Acad. Sci.*, **224**, 125.
———. 1974*a*, in *Gravitational Radiation and Gravitational Collapse*, ed. C. DeWitt-Morett (Dordrecht: Reidel).
———. 1974*b*, in *Confrontation of Cosmological Theories with Observational Data*, ed. M. S. Longair (Dordrecht: Reidel).
———. 1975, *Proc. 1974 Erice Meeting*, ed. de Sabbata (in press).
———. 1976*a*, in *Proc. 1975 Marcel Grossman Conference*, ed. R. Ruffini (in press).
———. 1976*b*, to appear.
Press, W. H., and Teukolsky, S. A. 1973, *Ap. J.*, **185**, 635.
Regge, T., and Wheeler, J. A. 1957, *Phys. Rev.*, **108**, 1063.
Robinson, D. 1975, *Phys. Rev. Letters*, **34**, 905.
Schmidt, B. J. 1971, *J. Gen. Rel. and Grav.*, **1**, 269.
———. 1972, *Comm. Math. Phys.*, **29**, 49.
———. 1974, *Comm. Math. Phys.*, **36**, 73.
Seifert, H.-J. 1971, *J. Gen. Rel. and Grav.*, **1**, 247.
Simpson, M., and Penrose, R. 1973, *Int. J. Theoret. Phys.*, **7**, 183.
Steinmüller, B., King, A. R., and Lasota, J. P. 1975, *Phys. Letters*, **51A**, 191.
Stewart, J. M. 1975, *Proc. Roy. Soc. London*, **A344**, 65.
Szekeres, P. 1975, preprint.
Teukolsky, S. A. 1973, *Ap. J.*, **185**, 635.
Tomimatsu, A., and Sato, H. 1972, *Phys. Rev. Letters*, **29**, 1344.
Vishveshwara, C. V. 1970, *Phys. Rev.*, **D1**, 2870.
Witten, L. 1974, in *Gravitational Radiation and Gravitational Collapse*, ed. C. DeWitt-Morette (Dordrecht: Reidel).
Yodzis, P., Seifert, H.-J., and Muller zum Hagen, H. 1973, *Comm. Math. Phys.*, **34**, 135.
———. 1974, *Comm. Math. Phys.*, **37**, 29.
Zel'dovich, Ya. B. 1973, *Soviet Phys.—JETP*, **37**, 33.

97
THE METRIC AND CURVATURE PROPERTIES OF $\mathcal{H}$-SPACE

Proc. Roy. Soc. Lond. Ser. A **363** (1978), no. 1715, 445–68.
With R. O. Hansen, E. T. Newman, and K. P. Tod

Newman's $\mathcal{H}$-space, its Ricci-flatness, its relation to asymptotic twistor space, and to the non-linear graviton construction of Chapters 67 and 70, are considered in some detail.

 The metric and curvature properties of $\mathcal{H}$-space. *Proc. Roy. Soc. Lond. Ser. A* **363** (1978), no. 1715, 445–68

The metric and curvature properties of $\mathscr{H}$-space†

BY R. O. HANSEN,[1,2]‡ E. T. NEWMAN[1], R. PENROSE[2], F.R.S., AND K. P. TOD[3]§

[1] *Department of Physics and Astronomy, University of Pittsburgh;*
[2] *Mathematical Institute, Oxford University;*
[3] *Department of Mathematics and Statistics, University of Pittsburgh*

(*Received* 14 *March* 1978)

The space $\mathscr{H}$ of asymptotically (left-) shear-free cuts of the $\mathscr{I}^+$ (good cuts) of an asymptotically flat space-time $\mathscr{M}$ is defined. The connection between this space and the asymptotic projective twistor space $P\mathscr{T}$ of $\mathscr{M}$ is discussed, and this relation is used to prove that $\mathscr{H}$ is four-complex-dimensional for sufficiently 'calm' gravitational radiation in $\mathscr{M}$. The metric on $\mathscr{H}$-space is defined by a simple contour integral expression and is found to be complex Riemannian. The good cut equation governing $\mathscr{H}$-space is solved to three orders by a Taylor series and the solution is used to demonstrate that the curvature of $\mathscr{H}$-space is always a self dual (left flat) solution of the Einstein vacuum equations.

1. INTRODUCTION

An isolated system, in general relativity, has a gravitational field which falls off towards infinity to give a space-time geometry which is asymptotically Minkowskian. If the system is stationary, the fall-off is sufficiently fast ($\sim r^{-3}$) that the Poincaré group may be recognized as an asymptotic symmetry group, and, consequently, well-defined concepts of total energy-momentum and total angular momentum unambiguously arise (Sachs 1962). However, for a non-stationary system from which there is outgoing gravitational radiation, the space-time curvature falls off rather more slowly ($\sim r^{-1}$) in future null directions, with the implication that the Poincaré group does not emerge in any very evident way as a symmetry of future null infinity, $\mathscr{I}^+$. Instead a much larger (infinite-parameter) group, the Bondi–Metzner–Sachs (B.M.S.) group has been introduced (Bondi *et al.* 1962, Sachs 1962) to describe the asymptotic symmetries. This is sufficient for the concept of energy-momentum (since there is a uniquely defined translation normal subgroup) but it does not give rise to an unambiguous concept of angular momentum.

† Work supported in part by a N.S.F. grant.
‡ Present address: Department of Mathematics, University of California, Berkeley.
§ Present address: Mathematical Institute, Oxford University.

For reasons of this sort it seems desirable to have an alternative approach, in the study of the asymptotic structure of an asymptotically flat space-time $\mathscr{M}$, to one based on the B.M.S. group. Such an alternative arises from a study of *complexified* future null infinity $\mathbb{C}\mathscr{I}^+$ and leads to the concept of $\mathscr{H}$-space introduced by Newman (1976) and explored in some of its detailed aspects by Newman and coworkers, (Ko *et al.* 1977). The approach relates closely to the theory of twistors (Penrose 1975) and provided the starting point for the study of asymptotic twistors (Penrose & MacCallum 1973, Penrose 1975) and the 'non-linear graviton' (Penrose 1976). In the present paper we explore the $\mathscr{H}$-space approach in some detail, setting up the correspondence with twistor theory and establishing some of the basic theorems.

The essential reason for the need for complexification of $\mathscr{I}^+$ lies in the fundamental problem of the B.M.S. approach: how are we to single out a canonical Poincaré subgroup of the B.M.S. group, with respect to which the angular momentum concept is to be defined? Such a canonical Poincaré subgroup would contain a 4-parameter system of Lorentz subgroups, one for each 'origin', the set of such 'origins' constituting an abstract 4-manifold $\mathbb{R}\mathscr{H}$ with a Minkowski space structure. Each of these Lorentz subgroups consists of the B.M.S. transformations leaving invariant a particular 'cut' of $\mathscr{I}^+$ (i.e. a smooth 2-sphere on $\mathscr{I}^+$ which meets each generator exactly once). This preferred set of cuts of $\mathscr{I}^+$ thus uniquely determines the particular choice of Poincaré subgroup, so it would seem desirable that a canonical set of cuts should be definable in terms of the local geometry of $\mathscr{I}^+$. In the case of Minkowski space, such *good cuts* arise as the intersections with $\mathscr{I}^+$ of the future light cones of finite points. They are easily recognizable at $\mathscr{I}^+$ by the fact that their asymptotic complex shear σ^0 vanishes. In the case of a stationary space-time, also, this definition suffices (the only essential difference being that $\mathbb{R}\mathscr{H}$ is not canonically identifiable with $\mathscr{M}$, in general, as it had been in the Minkowski case). However, when gravitational radiation is present, good cuts of $\mathscr{I}^+$ defined in this way do not exist. The reason is basically that, in the first instance, gravitational radiation will affect the value of σ^0 in an arbitrary way (Sachs 1962) and secondly σ^0 is complex so that two real degrees of freedom per point of the cut would be needed to make it vanish, whereas only one real degree of freedom per point is available along each generator of $\mathscr{I}^+$ in the choice of cut. But by replacing $\mathscr{I}^+$ by its complexification $\mathbb{C}\mathscr{I}^+$ this difficulty is removed. And, while it may seem at first that complexification is a high price to pay – particularly as, even then, a canonical Poincaré group does not arise at this stage – we are rewarded by the very remarkable geometric structure that the resulting $\mathscr{H}$-space possesses.

Each point of this space (which is sometimes referred to as 'heaven') is a complex cut of $\mathbb{C}\mathscr{I}^+$ for which $\sigma^0 = 0$. As an alternative, the choice $\tilde{\sigma}^0 = 0$ could be made where $\tilde{\sigma}^0$ is the analytic extension of $\bar{\sigma}^0$, the complex conjugate of σ^0. This yields the complex conjugate space, $\bar{\mathscr{H}}$. It can be rigorously shown (by invoking some theorems of Kodaira on deformations of complex manifolds, as we shall demonstrate in § 3) that with the assumption that the gravitational radiation is not too

violent, $\mathscr{H}$-space and $\tilde{\mathscr{H}}$-space are 4-dimensional complex manifolds. Furthermore, there is a natural definition of complex metric on $\mathscr{H}$ and $\tilde{\mathscr{H}}$ defined by a contour integral expression, and it turns out, remarkably, that $\mathscr{H}$ and $\tilde{\mathscr{H}}$ automatically satisfy the Einstein vacuum equations. This is irrespective of whether or not the original space-time is vacuum, provided that the appropriate asymptotic flatness conditions hold. The curvature of $\mathscr{H}$ turns out, in addition, to be self-dual (which can happen because $\mathscr{H}$ is a complex space) so $\mathscr{H}$ is what is called 'left-flat'. Correspondingly, $\tilde{\mathscr{H}}$ is 'right-flat'.

The space $\mathscr{H}$ can be given a striking interpretation in those circumstances when $\mathscr{H}$ itself may (in an appropriate sense) be regarded as asymptotically flat. The outgoing radiation field of the original space-time can be separated into a self-dual and an anti-self-dual part. The space $\mathscr{H}$ may then be regarded as the purely radiative (i.e. with no sources) space-time constructed from the self-dual radiation and in which the anti-self-dual radiation has been set to zero. Correspondingly, $\tilde{\mathscr{H}}$ consists of just the anti-self-dual radiation, with the self-dual part set to zero.

Though the construction of $\mathscr{H}$-space does not by itself lead to the reinstatement of the Poincaré group as an asymptotic symmetry group of $\mathscr{M}$ it may be regarded as a half-way stage towards doing so. In those circumstances when $\mathscr{H}$ is appropriately asymptotically flat, its own $\tilde{\mathscr{H}}$-space can be constructed and this, since $\mathscr{H}$ contains no anti-self-dual field, must be a flat space-time. A Poincaré group then does arise and can be apparently related to one obtained earlier by a different method (Newman & Penrose 1966).

In § 2, we define the $\mathscr{H}$ and $\tilde{\mathscr{H}}$-spaces of an asymptotically flat curved space-time, and derive the good cut equation which generates them. § 3 is devoted to a discussion of the connection between $\mathscr{H}$-space and asymptotic twistor space, and of the manifold structure of the $\mathscr{H}$-space. In § 4, we solve the Taylor series expansion of solutions to the good cut equation at the first three orders. These solutions are used in § 5 to show that $\mathscr{H}$-space has a complex Riemannian structure, while §§ 6 and 7 are devoted to finding the $\mathscr{H}$-space curvature tensor in terms of the asymptotic radiation field of the original physical space.

2. The definition of $\mathscr{H}$-space

We begin with a discussion of Minkowski space, M. The future light cones of finite points of M intersect the future null infinity $\mathscr{I}^+$ of M in a four-parameter set of cuts. These cuts are distinguished from all other cuts of $\mathscr{I}^+$ by being the intersections with $\mathscr{I}^+$ of asymptotically shear-free null hypersurfaces (in fact for M, everywhere shear-free). Such cuts are referred to as good cuts.

The metric structure of Minkowski space may be described in terms of $\mathscr{I}^+$ by a remarkable integral expression. This very same expression will serve to define the metric in the case of a curved space-time, as it depends only on a certain structure, called the *strong conformal geometry* (Penrose 1974), that $\mathscr{I}^+$ possesses even in the general case.

Essentially the expression is as follows: If Σ and Σ' are two neighbouring cuts of $\mathscr{I}^+$, then their separation at each point of Σ may be measured in terms of a null vector U^i pointing along the generator of $\mathscr{I}^+$ at that point. Now, with any conformal scaling for which the metric on $\mathscr{I}^+$ is finite, there is associated a particular null vector field n^i tangent to the generators. The strong conformal geometry of $\mathscr{I}^+$ may be described as $n^i n^j q_{kl}$, the (degenerate) induced metric tensor of $\mathscr{I}^+$ being q_{kl}, and it determines $n^i n^j \mathrm{d}S$, where $\mathrm{d}S$ is the element of surface area of cross-section of $\mathscr{I}^+$ (cf. Penrose 1974). We have $U^i = Un^i$ giving U as a function on Σ which measures the separation of Σ' from it. The invariance of strong conformal geometry thus assures us that the integral

$$\frac{1}{8\pi}\oint \frac{\mathrm{d}S}{U^2}$$

taken over the whole of Σ, must be invariant, i.e. independent of the choice of conformal factor. The integral exists provided U is nowhere vanishing – which is the case when the displacement from Σ to Σ' represents a timelike vector u^a in M. Its value turns out to be the reciprocal of $g_{ab}u^a u^b$ and thus serves to define the metric of M purely in terms of the structure of $\mathscr{I}^+$.

To investigate this in detail, it is convenient to introduce stereographic null coordinates in M and on $\mathscr{I}^+$ based on a timelike geodesic γ in M. Let t^a be the future-directed tangent vector to γ normalized by $t^a t_a = 2$, and let $u\sqrt{2}$ be proper time along γ. Introduce l^a as the null vector field tangent to the future light cones of points on γ normalized by $t^a l_a = 1$. The light cones are labelled by the value of u at the vertex. We label the generators of the null cones by the stereographic coordinates ζ and $\bar{\zeta}$, with†

$$\zeta = \mathrm{e}^{-\mathrm{i}\phi}\cot\tfrac{1}{2}\theta.$$

The fourth coordinate r is the affine parameter associated with l^a along null geodesics of the cones measured from γ. Now any point of M with position vector x^a in Cartesian coordinates is given stereographic null coordinates $(u_0, r_0, \zeta_0, \bar{\zeta}_0)$ by writing

$$x^{\mathbf{a}} = u_0 t^{\mathbf{a}} + r_0 l^{\mathbf{a}}(\zeta_0, \bar{\zeta}_0),$$

where $l^{\mathbf{a}}$ has been chosen to have Cartesian components

$$l^{\mathbf{a}} = \frac{1}{\sqrt{2}}\left(1, \frac{\zeta+\bar{\zeta}}{1+\zeta\bar{\zeta}}, \frac{\mathrm{i}(\zeta-\bar{\zeta})}{1+\zeta\bar{\zeta}}, \frac{-(1-\zeta\bar{\zeta})}{1+\zeta\bar{\zeta}}\right). \tag{2.1}$$

In the limit $r\to\infty$, $(u, \zeta, \bar{\zeta})$ are coordinates on $\mathscr{I}^+$. We express the light cone of a point $x^{\mathbf{a}}$ in the interior as

$$u = f(x^{\mathbf{a}}, r, \zeta, \bar{\zeta})$$

and as $r\to\infty$ the good cut, which is the intersection of this light cone with infinity is given by

$$u = Z(x^{\mathbf{a}}, \zeta, \bar{\zeta}) \equiv x^{\mathbf{a}} l_{\mathbf{a}}(\zeta, \bar{\zeta}). \tag{2.2}$$

† It should be noted that the definition of ζ used here coincides with that of $\bar{\zeta}$ in Newman & Penrose (1966). Thus their equations (3.7) and (3.8), which define $\eth$ and $\bar{\eth}$ in terms of $\partial/\partial\theta$ and $\partial/\partial\phi$, are modified by replacing i by $-\mathrm{i}$.

We recall that the operators $\eth$ and $\bar{\eth}$ acting on a function of spin-weight s are defined by (Newman & Penrose 1966)

$$\eth\chi = 2P^{1-s}\,\partial(P^s\chi)/\partial\zeta,$$

$$\bar{\eth}\chi = 2P^{1+s}\,\partial(P^{-s}\chi)/\partial\bar{\zeta},$$

where

$$P = \tfrac{1}{2}(1+\zeta\bar{\zeta})$$

and give functions of spin-weights $s+1$ and $s-1$ respectively. From (2.1) and (2.2) Z is spin-weight zero and consists of $l = 0$ and 1 spherical harmonics and

$$\eth^2 Z = 0. \tag{2.3}$$

A regular spin-weighted function χ should more properly be thought of as a not necessarily analytic global section of a complex line bundle over the Riemann sphere of ζ and is thus represented by a pair of functions $\chi(\zeta,\bar{\zeta})$ and $\hat{\chi}(\zeta^{-1},\bar{\zeta}^{-1})$ both of which are non-singular at finite values of their arguments. If χ has spin-weight s then

$$\hat{\chi}(\zeta^{-1},\bar{\zeta}^{-1}) = \zeta^s\bar{\zeta}^{-s}\chi(\zeta,\bar{\zeta}).$$

(When we complexify $\mathscr{I}^+$, $\bar{\zeta}$ is replaced in the above expressions by $\tilde{\zeta}$ and $\bar{\eth}$ becomes a corresponding operator $\tilde{\eth}$.)

Now, all solutions of (2.3) defined globally on the ζ-sphere are of the form (2.2). Thus equation (2.3) provides a convenient characterization of the good cuts. The space of good cuts, that is the space of globally defined solutions of (2.3) is defined to be the real $\mathscr{H}$-space, $\mathbb{R}\mathscr{H}$, of Minkowski space.

A tangent vector at a point of $\mathbb{R}\mathscr{H}$ may be described by considering a one-parameter family $x^a(\lambda)$ of points in M and the corresponding one-parameter family of good cuts, $Z(\lambda,\zeta,\bar{\zeta})$.

If $u^a = \mathrm{d}x^a/\mathrm{d}\lambda$ is the tangent to the curve in M then define

$$U = \mathrm{d}Z/\mathrm{d}\lambda = u^a l_a(\zeta,\bar{\zeta}), \tag{2.4}$$

where $U > 0$ if u^a is time-like. Thus a tangent vector in $\mathbb{R}\mathscr{H}$ is conveniently described by a function $U(\zeta,\bar{\zeta})$ as in (2.4). The metric of M is now given by

$$\eta_{ab}\,u^a u^b = 8\pi\left[\oint\frac{\mathrm{d}S}{U^2}\right]^{-1}, \tag{2.5}$$

where $\mathrm{d}S = 2\mathrm{d}\zeta\wedge\mathrm{d}\bar{\zeta}/\mathrm{i}(1+\zeta\bar{\zeta})^2$ is the element of surface area of the ζ-sphere, and for time-like u^a (in which case $U > 0$) the integration is over the whole sphere. If the vector u^a is spacelike the integration surface can be found in the complexification of this sphere, this surface being a continuous deformation of the real sphere. If however u^a becomes null the singularities of the integrand 'pinch' and no such deformation surface exists, so the integral in (2.5) becomes infinite. Equation (2.5) may be proved by substituting (2.4) for U and integrating. Thus the structure

of Minkowski space is completely recovered by this $\mathbb{R}\mathcal{H}$-space construction. Note that the same analysis can be carried out for complexified Minkowski space where (2.5) is regarded as an integral over a sphere deformed into the complex.

The remainder of this paper is concerned with the extension of these ideas to an asymptotically flat, curved space-time $\mathcal{M}$. In general, one would not expect to recover an exact description of the physical geometry from structures defined only on $\mathcal{I}^+$; indeed a curved $\mathcal{M}$ is never isometric to its $\mathcal{H}$-space.

We begin by generalizing the discussion leading to (2.3) to the case of asymptotically flat solutions to the vacuum Einstein (or Einstein–Maxwell) equations. In a null coordinate system $(u, r, \zeta, \bar{\zeta})$ (u is a Bondi-type retarded null coordinate, r an affine parameter along the associated null geodesics, and ζ and $\bar{\zeta}$ are complex stereographic coordinates which label the geodesics in the null surfaces) the equation describing an *arbitrary* null surface can be written in the form

$$u = f(\zeta, \bar{\zeta}, r). \tag{2.6}$$

The intersection of the surface (2.6) with $\mathcal{I}^+$ (a cut of $\mathcal{I}^+$) is again given by

$$u = \lim_{r \to \infty} f(\zeta, \bar{\zeta}, r) \equiv Z(\zeta, \bar{\zeta}). \tag{2.7}$$

If the cut function $u = Z(\zeta, \bar{\zeta})$ is known, the entire null surface can be reconstructed. We shall be concerned with the problem of finding null surfaces whose asymptotic shear vanishes.

Given the asymptotic shear† $\sigma^0(u, \zeta, \bar{\zeta})$ of the (Bondi) null cone u = constant it follows from the transformation law of σ^0 under supertranslations (Sachs 1962) that the asymptotic shear of the cone associated with the cut function $Z(\zeta, \bar{\zeta})$ is

$$\sigma'^0(\zeta, \bar{\zeta}) = \sigma^0(Z(\zeta, \bar{\zeta}), \zeta, \bar{\zeta}) - \eth^2 Z. \tag{2.8}$$

The condition for a cut to have vanishing asymptotic shear is thus

$$\eth^2 Z = \sigma^0(Z(\zeta, \bar{\zeta}), \zeta, \bar{\zeta}). \tag{2.9}$$

However, since σ^0 is complex and Z is real, (2.9) generally has no solutions: in general asymptotically flat space-times there exist no asymptotically shear-free null hypersurfaces.

The situation is very different if we consider an analytic extension of $\mathcal{M}$ in the neighbourhood of $\mathcal{I}^+$ into the complex domain. (By this we mean a complex manifold containing $\mathcal{M} \cup \mathcal{I}^+$ for which geometrical quantities defined on $\mathcal{M} \cup \mathcal{I}^+$ extend to holomorphic quantities. We thus require that $\sigma^0(u, \zeta, \bar{\zeta})$ be analytic in u and in the real and imaginary parts of ζ. The function σ^0 extends to a holomorphic function of independent complex coordinates u, ζ and $\tilde{\zeta}$ where, on $\mathcal{I}^+$, u takes real values and $\tilde{\zeta} = \bar{\zeta}$.) Now since u is complex, we can take *complex* cut functions $u = Z(\zeta, \tilde{\zeta})$; equation (2.9), rewritten as

$$\eth^2 Z(\zeta, \tilde{\zeta}) = \sigma^0(Z, \zeta, \tilde{\zeta}), \tag{2.10}$$

† The shear of the null rays lying in the Bondi cones has the form $\sigma(u, r, \zeta, \bar{\zeta}) = \sigma^0(u, \zeta, \bar{\zeta})\, r^{-2} + O(r^{-3})$, where σ^0, which is referred to as the asymptotic shear, determines the outgoing gravitational radiation field (cf. Newman & Penrose 1966).

becomes the equation for cut functions which yield asymptotically shear-free, *complex* null cones. We refer to (2.10) as the *good cut equation*, and to the solutions as *good cut functions*.

In § 3, we shall show that the regular solutions of (2.5) for appropriately 'calm' σ^0, form a four complex-parameter set i.e. that the good cuts form a four-dimensional manifold $\mathscr{H}$. Thus, we can write the general good cut function as

$$u = Z(z^{\mathfrak{a}}, \zeta, \tilde{\zeta}), \tag{2.11}$$

where the $z^{\mathfrak{a}}$ are four complex parameters (or coordinates) which label the good cuts (i.e. points of $\mathscr{H}$-space) and where Z is holomorphic in all six variables.

Note that the asymptotic shear σ^0 is complex-valued even on real null infinity; thus $\overline{\sigma}^0$, the complex conjugate of σ^0, defines by analytic extension an *independent* field $\tilde{\sigma}^0$ on complex null infinity. We could equally well have chosen cuts $u = \tilde{Z}(\zeta, \tilde{\zeta})$ such that $\tilde{\sigma}'^0 = 0$. Clearly, these too will have to form a four-dimensional manifold. We refer to the manifold of cuts with $\tilde{\sigma}'^0 = 0$ as an $\tilde{\mathscr{H}}$-space. We shall consider only $\mathscr{H}$-spaces; the arguments presented here will hold for $\tilde{\mathscr{H}}$-spaces as well. For flat or stationary space-times, $\mathscr{H}$ and $\tilde{\mathscr{H}}$ coincide as complex Minkowski space.

The affine distance along the generators (ζ and $\tilde{\zeta}$ = const.) of $\mathbb{C}\mathscr{I}^+$ between two neighbouring good cuts (or points of $\mathscr{H}$-space) with coordinates $z^{\mathfrak{a}}$ and $z^{\mathfrak{a}} + \mathrm{d}z^{\mathfrak{a}}$ is given by

$$\mathrm{d}Z = Z_{,\mathfrak{a}}\, \mathrm{d}z^{\mathfrak{a}}. \tag{2.12}$$

Let λ be a parameter along the curve through $z^{\mathfrak{a}}$ and $z^{\mathfrak{a}} + \mathrm{d}z^{\mathfrak{a}}$, then we have

$$\frac{\mathrm{d}Z}{\mathrm{d}\lambda} = Z_{,\mathfrak{a}} \frac{\mathrm{d}z^{\mathfrak{a}}}{\mathrm{d}\lambda}. \tag{2.13}$$

The expression on the left of (2.13) represents the infinitesimal supertranslation which maps the good cut with coordinates $z^{\mathfrak{a}}$ to the good cut with coordinates $z^{\mathfrak{a}} + \mathrm{d}z^{\mathfrak{a}}$. The term $\mathrm{d}z^{\mathfrak{a}}/\mathrm{d}\lambda$ on the right of (2.13) defines a tangent vector $u^{\mathfrak{a}}$ at the point $z^{\mathfrak{a}}$ in $\mathscr{H}$-space. Letting

$$U(z^{\mathfrak{a}}, \zeta, \bar{\zeta}) = Z_{,\mathfrak{a}}\, u^{\mathfrak{a}} \tag{2.14}$$

we define a norm $\|u\|^2$ on the tangent vectors $u^{\mathfrak{a}}$ of $\mathscr{H}$-space by

$$\|u\|^2 = 8\pi \left(\oint \frac{\mathrm{d}S}{U^2} \right)^{-1}, \tag{2.15}$$

where now $\mathrm{d}S = 2\mathrm{d}\zeta \wedge \mathrm{d}\tilde{\zeta}/\mathrm{i}(1 + \zeta\tilde{\zeta})^2$ is the area element on the complex unit sphere, and the integral is taken over the real sphere, $\tilde{\zeta} = \bar{\zeta}$ (or, in order to avoid singularities, over an appropriate deformation of the real sphere into the complex). We establish in § 5 the remarkable fact that this norm induces a *complex Riemannian* metric on $\mathscr{H}$-space:

$$\|u\|^2 = g_{\mathfrak{a}\mathfrak{b}}\, u^{\mathfrak{a}} u^{\mathfrak{b}} \tag{2.16}$$

where $g_{\mathfrak{a}\mathfrak{b}}$ is a complex, non-degenerate, symmetric tensor field on $\mathscr{H}$.

In this section we examine the relation of $\mathscr{H}$-space to deformed twistor space. Some of the details of this construction have been given elsewhere (Penrose 1976). Here we concentrate on some different aspects of the construction. In particular, we shall show how two theorems of Kodaira (1962, 1963), concerning the stability of compact holomorphic submanifolds of a complex manifold under deformations of its complex structure, may be used to obtain the result that the good cut equation (2.10) has the required four-parameter set of solutions provided that σ^0 is sufficiently 'calm.' We shall also show how the metric (2.15) of $\mathscr{H}$-space can be written in a form not involving a contour integral.

Detailed accounts of twistor theory may be found in Penrose & MacCallum (1973), Penrose (1975). For our present purposes we require only a small portion of this theory. We are here concerned with deformations of dual twistor space T^* rather than the original space T, for the notational reason that we have chosen (2.10) as the good cut equation rather than its complex conjugate (i.e. we are choosing σ^0 to vanish rather than $\tilde{\sigma}^0$. This gives $\mathscr{H}$ as left-flat rather than right-flat (Penrose 1976).) Of course, as mentioned earlier, all the complex conjugate results will apply equally well with σ^0 replaced by $\tilde{\sigma}^0$, left flat by right flat and T^* by T.

A flat-space twistor V^α may be represented up to proportionality as a totally null 2-plane $\mathbb{C}V$ in complex Minkowski space $\mathbb{C}M$ (Penrose 1975). This description arises as follows. The twistor V^α is defined by its spinor parts $(\omega^A, \pi_{A'})$. The point $x \in \mathbb{C}M$, with position vector $x^{AA'}$ (in 2-spinor notation) lies on $\mathbb{C}V$ whenever

$$\omega^A = \mathrm{i} x^{AA'} \pi_{A'}. \tag{3.1}$$

If y^a is another such position vector satisfying the same relation, it follows that

$$x^{AA'} - y^{AA'} = \lambda^A \pi^{A'} \tag{3.2}$$

for some λ^A. All such differences, $x^a - y^a$, must be null and orthogonal to one another ($\pi^{A'}$ is fixed, but λ^A varies) showing that $\mathbb{C}V$ is indeed totally null. (For this it must be assumed that $\pi_{A'} \neq 0$; otherwise – and then assuming that ω^A does not also vanish – $\mathbb{C}V$ is at infinity.) Similarly, a dual flat-space twistor W_α is also represented up to proportionality by a totally null 2-plane $\mathbb{C}W$, but of an opposite kind. For if $(\lambda_A, \mu^{A'})$ are the spinor parts of W_α we have (assuming $\lambda_A \neq 0$)

$$\mu^{A'} = -\mathrm{i} x^{AA'} \lambda_A. \tag{3.3}$$

If x^a lies on $\mathbb{C}W$, and if y^a also lies on $\mathbb{C}W$ we still have (3.2) holding, but now it is λ^A that is fixed and $\pi^{A'}$ that varies. (We need $\lambda_A \neq 0$, for finite $\mathbb{C}W$.)

The 2-planes $\mathbb{C}V$ and $\mathbb{C}W$ meet $\mathbb{C}\mathscr{I}^+$ in complex null geodesics V and W, lying, respectively, in $\zeta = \text{const.}$ and in $\tilde{\zeta} = \text{const.}$; and these geodesics define $\mathbb{C}V$ and $\mathbb{C}W$ uniquely, i.e. they define the twistors V^α and W_α up to proportionality. With

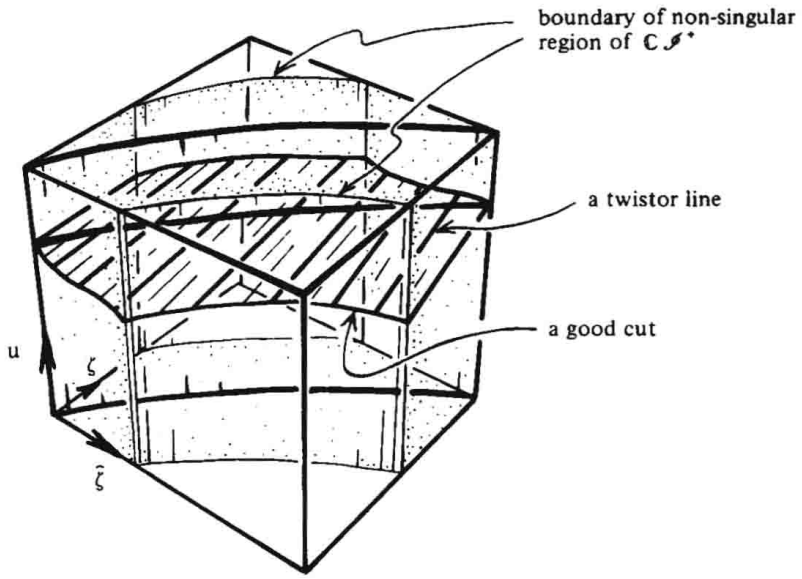


FIGURE 1. $\mathbb{C}\mathscr{I}^+$ showing a good cut ruled by twistor lines.

the coordinates $(u, \zeta, \tilde{\zeta})$ of § 2 we find (Hansen & Newman 1975) that V and W can be represented as:

$$u = V(\tilde{\zeta});\ \zeta = \zeta_0,\ \tilde{\partial}^2 V = 0 \tag{3.4}$$

and as

$$u = W(\zeta);\ \tilde{\zeta} = \tilde{\zeta}_0,\ \partial^2 W = 0 \tag{3.5}$$

respectively.

In a curved asymptotically flat (analytic) space-time, $\mathscr{M}$, we can employ the concept of asymptotic twistor (Penrose & MacCallum 1973, Penrose 1975). An asymptotic twistor or dual twistor (except those 'lying entirely on $\mathscr{I}^+$' which would correspond to $\pi_{A'} = 0$ or $\lambda_A = 0$) can again be represented up to proportionality on $\mathbb{C}\mathscr{I}^+$ as complex null geodesics V and W, just as before, but where now, using the Bondi-type coordinates of § 2, we have the respective equations

$$u = V(\tilde{\zeta});\quad \zeta = \zeta_0,\quad \tilde{\partial}^2 V = \tilde{\sigma}^0 \tag{3.6}$$

and

$$u = W(\zeta);\quad \tilde{\zeta} = \tilde{\zeta}_0,\quad \partial^2 W = \sigma^0. \tag{3.7}$$

It is a simple matter, following from the definition of σ^0, to see that the twistor lines are in fact null geodesics of $\mathbb{C}\mathscr{I}^+$. Comparison with (2.10) shows that the two types of good cuts of $\mathbb{C}\mathscr{I}^+$, yielding points of $\tilde{\mathscr{H}}$ and of $\mathscr{H}$, are ruled by twistor lines V and by dual twistor lines W, respectively.

Let us focus attention on $\mathscr{H}$. A good cut representing a single point of $\mathscr{H}$ is depicted in figure 1. It is a ruled surface generated by null geodesics in each $\tilde{\zeta}$ = constant plane. However, only a certain limited range of the complex variables $u, \zeta, \tilde{\zeta}$, can be expected to describe non-singular points of $\mathbb{C}\mathscr{I}^+$. All we know is that $\mathscr{I}^+$ is non-singular and analytic. This means that some complex neighbourhood of $\mathscr{I}^+$ (i.e. of $u = \bar{u}$, $\tilde{\zeta} = \tilde{\zeta}$), is non-singular. We call such a neighbourhood a *thickening*

of $\mathscr{I}^+$ into the complex and label this neighbourhood $\mathbb{C}\mathscr{I}^+$. When the situation is close to that of Minkowski space we would expect a large radius of convergence about the points of $\mathscr{I}^+$ of the geometric quantities such as σ^0, and consequently, we would obtain a large thickening, $\mathbb{C}\mathscr{I}^+$. In this case we say that σ^0 is 'calm'. On the other hand, we could have a comparatively *rough* but still analytic σ^0 for which the thickening of $\mathscr{I}^+$ would be narrow. In either case we could envisage good cuts of $\mathbb{C}\mathscr{I}^+$ for which σ'^0 vanishes. But for the best chance of such cuts existing globally (as they must, in order to qualify strictly as 'cuts') we should choose our thickening to be as wide as possible in the complex parameter range of u away from $u = \bar{u}$ but very narrow away from $\tilde{\zeta} = \bar{\zeta}$. This will still ensure that for each value of $\tilde{\zeta}$ (including $\tilde{\zeta} = \infty$) we have a portion of a twistor line in the cut, provided the cut exists, and the wider the range in complex u that is available, the more chance there is that the cut should exist.

Note that the differential equation in (3.7) is second-order in the one variable ζ, so (as is to be expected for a geodesic) a dual twistor line is fixed once a point on it is fixed together with a tangent direction there. Given any $\tilde{\zeta}$, we thus have a two-complex-parameter family of dual twistor lines. So allowing $\tilde{\zeta}$ to vary, we have three complex parameters in all. The space of these twistor lines is thus a space of three complex dimensions – the space $P\mathscr{T}^*$ of *projective asymptotic dual twistors*. Indeed, this construction gives $P\mathscr{T}^*$ as a complex three-manifold because the dual twistor lines are defined in $\mathbb{C}\mathscr{I}^+$ by the entirely holomorphic equations (3.7) in the independent complex variables u, ζ, $\tilde{\zeta}$. Also there need be no difficulty about the segments of dual twistor lines possibly intersecting the relevant region of $\mathbb{C}\mathscr{I}^+$ more than once. In effect we may take this region to be an infinitesimally narrow neighbourhood of $\bar{\zeta} = \tilde{\zeta}$ in the $(\zeta, \tilde{\zeta})$-space, each line being determined on $\bar{\zeta} = \tilde{\zeta}_0$.

Let us next see how a point of $\mathscr{H}$ is to appear in terms of $P\mathscr{T}^*$. What we require is the whole one-parameter holomorphic family of dual twistor lines that is given as $\tilde{\zeta}_0$ varies along a good cut (the cut function $Z(\zeta, \tilde{\zeta})$ being holomorphic also in $\tilde{\zeta}$). This family corresponds to a holomorphic curve in $P\mathscr{T}^*$. Furthermore, this curve is compact and has the topology of S^2, being in 1 : 1 holomorphic correspondence with the $\tilde{\zeta}$-sphere. Our problem of finding solutions to the good cut equation (2.3) thus reduces to one of finding compact holomorphic curves in the complex manifold $P\mathscr{T}^*$.

The existence of such suitable curves, for sufficiently calm σ^0 functions, is in fact ensured by a theorem of Kodaira (1963). We may think of $P\mathscr{T}^*$ as arising as a complex deformation of *flat* dual projective twistor space PT^*, for which case the required holomorphic curves certainly exist – corresponding to the points of complex Minkowski space $\mathbb{C}M$, i.e. to the good cuts (2.2). In fact, these holomorphic curves are simply the projective lines in the complex projective three-space ($\mathbb{C}P^3$) that is PT. We now envisage the σ^0-function as belonging to a holomorphic family of such functions, parameterized by a further complex number τ varying over some neighbourhood Γ of the origin in $\mathbb{C}$, where $\tau = 0$ yields $\sigma^0 = 0$. We thus have a

holomorphic family of complex manifolds $P\mathcal{T}^*(\tau)$ for which $P\mathcal{T}^*(0)$ is a portion of the flat projective twistor space PT^*. If desired, we can choose σ^0 simply to be proportional to τ, this being written explicitly as

$$\sigma^0(u, \zeta, \tilde{\zeta}; \tau) = \tau\Sigma(u, \zeta, \tilde{\zeta}), \tag{3.8}$$

where the function Σ is holomorphic in some complex open neighbourhood Δ, in $\mathbb{C}\mathscr{I}^+$, of the region $u = \bar{u}$, $\tilde{\zeta} = \bar{\zeta}$.

Now when $\tau = 0$ we certainly have a four-complex parameter family of suitable compact holomorphic curves, namely those given by the complex Minkowskian good cuts which are close enough to the real Minkowskian good cuts that they are defined in the region Δ. Each such cut is given by an equation

$$u = \frac{(z^0 + z^3) - (z^1 + \mathrm{i}z^2)\,\zeta - (z^1 - \mathrm{i}z^2)\,\tilde{\zeta} + (z^0 - z^3)\,\zeta\tilde{\zeta}}{\sqrt{2}(1 + \zeta\tilde{\zeta})} \tag{3.9}$$

for fixed complex numbers z^0, z^1, z^2, z^3 which are adequately close to being real numbers. These curves are projective lines in PT^* which lie in the region $P\mathcal{T}^*(0)$, i.e. which are sufficiently close to PN^*, where PN^* is the system of flat-space twistor lines containing real points, i.e. which meet $\mathscr{I}^+$.

Select one of these compact holomorphic curves $Z \in P\mathcal{T}^*(0)$. We require to know if an analogue of Z still exists when τ is made non-zero. More precisely: does there exist a holomorphic family of compact holomorphic curves $Z(\tau)$, parameterized by τ, where $Z(\tau) \in P\mathcal{T}^*(\tau)$ for each $\tau \in \Gamma$? The concept 'holomorphic family' can be taken to mean that the curves $Z(\tau)$ trace out a complex two-manifold as τ varies, this two-manifold lying inside a certain complex four-manifold on which τ is a parameter ranging over Γ, and with each τ = const. defining $P\mathcal{T}^*(\tau)$. (This four-manifold is easily constructed. We first introduce τ as an extra coordinate, which, along with u, ζ, $\tilde{\zeta}$, serves to define a preliminary four-manifold with the complex structure of $\mathbb{C}\mathscr{I}^+ \times \Gamma$ on which $\sigma^0(u, \zeta, \tilde{\zeta}; \tau)$ is defined. The space of all dual twistor lines in this gives us the required four-manifold.) Now Kodaira's (1963) theorem assures us that this holomorphic family of curves $Z(\tau)$ does exist provided a certain condition is satisfied concerning $Z = Z(0)$ and its immediate neighbourhood in $P\mathcal{T}^*(0)$. This condition is that the first cohomology group of Z with coefficients in the sheaf $\mathcal{N}$ of germs of holomorphic cross-sections of the normal bundle of Z in $P\mathcal{T}^*(0)$ vanish; briefly

$$H^1(Z, \mathcal{N}) = 0. \tag{3.10}$$

For the meanings of these terms, the reader is referred to suitable reference works (e.g. Morrow & Kodaira 1971, Gunning 1966, Wells 1973). The essential point is that all we need examine is the perfectly standard situation of the immediate neighbourhood of a complex projective line $\mathbb{C}P^1$ in a complex projective three-space $\mathbb{C}P^3$ (i.e. of Z in $P\mathcal{T}^*(0)$), and the existence of a corresponding $Z(\tau)$ in $P\mathcal{T}^*(\tau)$ may be inferred.

The fact that (3.10) actually holds may be deduced in various ways. One of these† is to express $H^1(Z, \mathcal{N})$ as a direct sum

$$H^1(Z, \mathcal{N}) = H^1(Z, \mathcal{N}_1) \oplus H^1(Z, \mathcal{N}_2),$$

where $\mathcal{N}_1$ and $\mathcal{N}_2$ are the sheaves of germs of holomorphic sections of the normal bundle of Z in two distinct planes π_1, π_2 through Z, in PT'^*; then to invoke, from the theory of holomorphic line bundles over $\mathbb{C}P^1$ (cf. Gunning 1966), the fact that at most one of $H^0(Z, \mathcal{N}_i)$, $H^1(Z, \mathcal{N}_i)$ can be non-vanishing (for each i), finally observing that $H^0(Z, \mathcal{N}_i) \neq 0$, owing to the fact that neighbouring lines to Z in the plane π_i define cross-sections of the relevant normal bundle. In fact, $H^0(Z, \mathcal{N}_i) \cong \mathbb{C}^2$.

This establishes that there is at least *one* solution of the good cut equation (3.7) which is global over the $\tilde{\zeta}$-sphere, provided that $\tau \in \Gamma$ (with small enough Γ). However, we require rather more than this, namely a four-complex-dimensional manifold of such solutions. One way of rendering this conclusion highly plausible is to imagine the above argument to be repeated for each Z of the four-complex parameter family (3.9). More satisfactory, however, is to use another of Kodaira's theorems (1962) as applied to a *particular* $Z(\tau)$ (for fixed $\tau \in \Gamma$). This theorem states that a sufficient condition for $Z(\tau)$ to belong to an n-dimensional family of compact holomorphic curves lying in $P\mathcal{T}^*(\tau)$ is $H^1(Z(\tau), \mathcal{N}(\tau)) = 0$, where $\mathcal{N}(\tau)$ is the sheaf of germs of holomorphic cross-sections of the normal bundle of $Z(\tau)$ in $P\mathcal{T}^*(\tau)$. The dimension n of this family is then $n = \dim H^0(Z(\tau), \mathcal{N}(\tau))$. Now provided that $\tau \in \Gamma'$, for some small enough open neighbourhood $\Gamma' \subset \Gamma$ of the origin in $\mathbb{C}$, the normal bundle of $Z(\tau)$ in $P\mathcal{T}^*(\tau)$ must be analytically identical to the normal bundle of Z in PT'^*. This is because the latter bundle is what is called a *stable* $\mathbb{C}^2$ bundle over $\mathbb{C}P^1$. One way of showing this stability is to use the fact that the normal bundle of Z in PT'^* has the essential structure of a quadric surface which, by a theorem of Brieskorn (1964) is a stable complex manifold (cf. Morrow & Kodaira 1971). It follows that $H^1(Z(\tau), \mathcal{N}(\tau))$ and $H^0(Z(\tau), \mathcal{N}(\tau))$ must be constant in τ, if $\tau \in \Gamma'$. Kodaira's (1962) theorem tells us that since $H^1(Z(\tau), \mathcal{N}(\tau)) = 0$, the space $H^0(Z(\tau), \mathcal{N}(\tau))$ is actually the tangent space to the family of holomorphic curves in $P\mathcal{T}^*(\tau)$ near $Z(\tau)$. But the above arguments give

$$\begin{aligned} H^0(Z(\tau), \mathcal{N}(\tau)) &\cong H^0(Z, \mathcal{N}) \\ &\cong H^0(Z, \mathcal{N}_1) \oplus H^0(Z, \mathcal{N}_2) \\ &\cong \mathbb{C}^2 \oplus \mathbb{C}^2 \cong \mathbb{C}^4 \end{aligned} \tag{3.11}$$

so that $n = 4$ and the required space of holomorphic curves, i.e. of solutions of the good cut equation, is four-complex-dimensional as required, provided $\tau \in \Gamma'$ – which is the '*sufficiently calm*' condition.

† We are deeply indebted to M. F. Atiyah, F.R.S. for providing essentially this argument establishing (3.10), and also for acquainting one of us (R. P.) with the relevant ideas of sheaf cohomology theory and with Kodaira's theorem and the basic notions concerning stable and unstable bundles.

This establishes that $\mathcal{H}$-space (and hence also $\mathcal{H}^{\sim}$-space) is a four-dimensional complex manifold, for sufficiently calm gravitational radiation fields and for good cuts chosen sufficiently close to the Minkowskian ones.

4. The Taylor series

We have seen in the previous sections that the first derivative of the good cut function, i.e. $Z_{,\mathbf{a}}$, is related to tangent vectors; correspondingly the second and third derivatives are related to the connection and curvature tensor. Thus we construct the first three terms of a Taylor series solution to the good cut equation.

Using the freedom to supertranslate, we assume that the complex Bondi cut $u = 0$ is a good cut, in which case $Z = 0$ is a solution of the good cut equation ($\sigma^0(0, \zeta, \tilde{\zeta}) = 0$). This cut (or point of $\mathcal{H}$-space) may be chosen as the origin $z^{\mathbf{a}} = 0$ of a coordinate system; the general cut is expressed as a Taylor series about $z^{\mathbf{a}} = 0$, in the form

$$Z(z^{\mathbf{a}}, \zeta, \tilde{\zeta}) = Z^0_{\mathbf{a}} z^{\mathbf{a}} + \tfrac{1}{2} Z^0_{\mathbf{ab}} z^{\mathbf{a}} z^{\mathbf{b}} + \tfrac{1}{6} Z^0_{\mathbf{abc}} z^{\mathbf{a}} z^{\mathbf{b}} z^{\mathbf{c}} + \dots, \tag{4.1}$$

where $Z^0_{\mathbf{a}}, Z^0_{\mathbf{ab}}, \dots$, are the partial derivatives of Z with respect to $z^{\mathbf{a}}$ at the origin. This follows from the analyticity of σ^0, and hence of Z. Substituting (4.1) into (2.10), and equating the coefficients of powers of $z^{\mathbf{a}}$, we obtain the basic equations for the determination of the Taylor series:

$$\eth^2 Z^0_{\mathbf{a}} = \dot{\sigma}^0 Z^0_{\mathbf{a}}, \tag{4.2a}$$

$$\eth Z^0_{\mathbf{ab}} = \dot{\sigma}^0 Z^0_{\mathbf{ab}} + \ddot{\sigma}^0 Z_{\mathbf{a}} Z_{\mathbf{b}}, \tag{4.2b}$$

$$\eth^2 Z^0_{\mathbf{abc}} = \dot{\sigma}^0 Z^0_{\mathbf{abc}} + 3\ddot{\sigma}^0 Z^0_{(\mathbf{a}} Z^0_{\mathbf{bc})} + \dddot{\sigma}^0 Z^0_{\mathbf{a}} Z^0_{\mathbf{b}} Z^0_{\mathbf{c}}, \tag{4.2c}$$

$$\eth^2 Z^0_{\mathbf{a}_1 \dots \mathbf{a}_n} = \dot{\sigma}^0 Z^0_{\mathbf{a}_1 \dots \mathbf{a}_n} + J_n(Z^0_{\mathbf{a}_1}, \dots, Z^0_{\mathbf{a}_1 \dots \mathbf{a}_{n-1}}, \ddot{\sigma}^0, \dots, \sigma^{(n)0}), \tag{4.2d}$$

where $\dot{\sigma}^0, \dots, \sigma^{(n)0}$ are successive u-derivatives of σ^0 evaluated at $u = 0$. Note that the nth equation of (4.2) is linear in $Z^0_{\mathbf{a}_1 \dots \mathbf{a}_n}(\zeta, \tilde{\zeta})$; (4.2$a$) is homogeneous, the rest inhomogeneous, the inhomogeneous parts depending on the solutions to the previous equations. The homogeneous equation corresponding to (4.2)

$$\eth^2 Z^{\mathrm{H}}_{\mathbf{a} \dots \mathbf{c}} = \dot{\sigma}^0 Z^{\mathrm{H}}_{\mathbf{a} \dots \mathbf{c}} \tag{4.3}$$

is essentially (4.2a); thus, to any solution $Z^0_{\mathbf{a} \dots \mathbf{c}}$ of (4.2d) a solution of the homogeneous equation may be added, of the form $\gamma^{\mathbf{m}}_{\mathbf{a} \dots \mathbf{c}} Z^0_{\mathbf{m}}$.

So far, we have not imposed any restrictions on the coordinate system, except the requirement that $z^{\mathbf{a}} = 0$ correspond to the cut $Z = 0$ about which the expansion is made. We claim that, by a transformation of the form

$$\begin{aligned} z'^{\mathbf{a}} &= z'^{\mathbf{a}}(z^{\mathbf{b}}) \\ &= a^{\mathbf{a}}_{\mathbf{b}} z^{\mathbf{b}} + \tfrac{1}{2} a^{\mathbf{a}}_{\mathbf{bc}} z^{\mathbf{b}} z^{\mathbf{c}} + \tfrac{1}{6} a^{\mathbf{a}}_{\mathbf{bcd}} z^{\mathbf{b}} z^{\mathbf{c}} z^{\mathbf{d}} + \dots \end{aligned} \tag{4.4}$$

any solution of the homogeneous equation (4.3) may be added to (4.2d). In fact, the freedom expressed in (4.3) coincides with the freedom (4.4) in the choice of coordinate system. Care must be exercised in using the coordinate freedom (4.4).

Transformation of any of (4.2) results in changes of the inhomogeneous terms in every subsequent equation, which must then be solved anew.

The quantities in which we are interested are all tensorial; we may therefore use the coordinate freedom (4.4) to simplify the solution of equations (4.2). It turns out that our choices will lead, to third order, to Riemann normal coordinates about the origin. In § 6, we outline a technique which enables us to find the quantities of interest in a coordinate-free way; the results are, of course, equivalent.

We proceed to solve the Taylor series equations (4.2) beginning with (4.2a). Let V_0 be any scalar solution of (4.2a) (i.e. $\eth^2 V_0 = \dot{\sigma}^0 V_0$) which is regular and has no zeros for $\zeta = \bar{\zeta}$. The existence of such solutions is guaranteed by the arguments in § 3. Then by a constant rescaling, $V_0 \to cV_0$, we can set

$$V_0^{-2} = 1 + \eth W_0, \tag{4.5}$$

where W_0 is a (unique) $s = -1$ function of ζ and $\tilde{\zeta}$. If we now make the ansatz

$$Z^0_{\mathbf{a}} = V_0 l^*_{\mathbf{a}} \tag{4.6}$$

with

$$l^*_{\mathbf{a}} = l_{\mathbf{a}} + W_0 m_{\mathbf{a}}, \quad m_{\mathbf{a}} = \eth l_{\mathbf{a}}, \tag{4.7}$$

(4.2a) becomes

$$\eth^2 l_{\mathbf{a}} = 0. \tag{4.8}$$

It is convenient to choose the solution in the form

$$l_{\mathbf{a}} = \frac{1}{\sqrt{2}}\left(1, \frac{-(\zeta+\tilde{\zeta})}{1+\zeta\tilde{\zeta}}, \frac{-\mathrm{i}(\zeta-\tilde{\zeta})}{1+\zeta\tilde{\zeta}}, \frac{1-\zeta\tilde{\zeta}}{1+\zeta\tilde{\zeta}}\right) \tag{4.9}$$

though any other set $l'_{\mathbf{a}} = \gamma^{\mathbf{b}}_{\mathbf{a}} l_{\mathbf{b}}$, $\det(\gamma^{\mathbf{a}}_{\mathbf{b}}) \neq 0$ would do just as well. A choice different from ours is equivalent to a first order coordinate transformation (4.4). It will be seen later that (4.9) leads to Riemann normal coordinates.

To solve the second-order equations we make the ansatz

$$Z^0_{\mathbf{ab}} = V_0(A_2 l^*_{\mathbf{a}} l^*_{\mathbf{b}} + 2B_2 l^*_{(\mathbf{a}} m_{\mathbf{b})} + C_2 m_{\mathbf{a}} m_{\mathbf{b}}), \tag{4.10}$$

where A_2, B_2 and C_2 are coefficients depending on ζ and $\tilde{\zeta}$ to be determined from (4.2b). By substituting (4.10) into (4.2b) a system of equations is obtained whose first integrals (after some effort) are found to be

$$\eth(V_0^2 \eth A_2) = V_0^3 \ddot{\sigma}^0, \tag{4.11a}$$

$$\eth B_2 = -2V_0^{-2} A_2, \tag{4.11b}$$

$$\eth C_2 = -3V_0^{-2} B_2. \tag{4.11c}$$

Equation (4.11a) can be integrated explicitly by decomposing V_0 and $\ddot{\sigma}^0$ in spherical harmonics; then integrals of (4.11b) and (4.11c) may be obtained. We shall not give the details here as the result is not needed in what follows. Note that the system has a unique solution. Two free functions, one consisting of $s = 1$, $l = 1$ harmonics, and the other of the $s = 0$, $l = 0$ harmonic, arise in the integration of (4.11a). The $s = 0$, $l = 0$ function is fixed by the integrability condition of (4.11b)

(essentially that B_2, having negative spin-weight, cannot contain $l = 0$ harmonics). The $s = 1, l = 1$ function is similarly determined by the integrability conditions for (4.11*c*). Note further that all three coefficients vanish if $\dot{\sigma}^0 = 0$.

The uniqueness of the solution to (4.11) suggests that, instead of $\ddot{\sigma}_0$, C_2 may be regarded as the free datum in that system; (4.11*c*) then determines B_2, (4.11*b*) determines A_2, and (4.11*a*) determines $\ddot{\sigma}^0$.

We now turn to the third-order coefficients Z^0_{abc}. Define

$$\left.\begin{aligned} n_a &= l_a + \eth\tilde{\eth} l_a, \\ \tilde{m}^*_a &= \tilde{\eth} l_a + W_0 n_a, \end{aligned}\right\} \tag{4.12}$$

and make the ansatz

$$\begin{aligned} Z^0_{abc} = V_0 [&A_3 l^*_a l^*_b l^*_c + 3B_3 l^*_{(a} l^*_b m_{c)} + 3C_3 l^*_{(a} m^*_b m_{c)} \\ &+ D_3 m_a m_b m_c + 3E_3 l^*_{(a} l^*_b \tilde{m}^*_{c)} + 3F_3 (l^*_{(a} l^*_b n_{c)} \\ &+ 2 l^*_{(a} m_b \tilde{m}^*_{c)}) + 3G_3 (2 l^*_{(a} m_b n_{c)} + m_{(a} m_b \tilde{m}^*_{c)}) \\ &+ 3H_3 m_{(a} m_b n_{c)}]. \end{aligned} \tag{4.13}$$

(We emphasize that this is not obvious and that a great deal of trial and error was involved before this ansatz was made.)

By substituting (4.13) into (4.2*c*) we eventually find

$$\left.\begin{aligned} \eth E_3 &= V_0^{-2} \mathscr{E}_3; \ \eth F_3 = -2 V_0^{-2} E_3, \\ \eth G_3 &= -3 V_0^{-2} F_3; \ \eth H_3 = -4 V_0^{-2} G_3, \end{aligned}\right\} \tag{4.14}$$

where $\mathscr{E}_3$ is a function of integration which is, at the moment, an arbitrary combination of $s = 2, l = 2$ spherical harmonics and

$$\eth B_3 = -2 V_0^{-2} A_3 + 3F_3 + V_0^{-2} (V_0^2 B_2 \eth A_2 + 2A_2^2), \tag{4.15a}$$

$$\eth C_3 = -3 V_0^{-2} B_3 + 3G_3 + V_0^{-2} (V_0^2 C_2 \eth A_2 + A_2 B_2), \tag{4.15b}$$

$$\eth D_3 = -4 V_0^{-2} C_3 + 3H_3 + 2 V_0^{-2} (B_2^2 - A_2 C_2), \tag{4.15c}$$

$$\eth [V_0^2 (\eth A_3 - 3E_3)] - 3 V_0^2 \eth E_3 = V_0^4 \ddot{\sigma}^0 + 3A_2 \eth (V_0^2 \eth A_2). \tag{4.16}$$

With an assumed $\mathscr{E}_3$, (4.14) can be integrated (in the form of a spherical harmonic decomposition) the integration constants always being determined by the requirement of regularity of the functions. In the same manner (4.16) and then (4.15*a*) and (4.15*b*) can be uniquely integrated. It is finally the requirement that D_3, in (4.15*c*), be regular that determines the $\mathscr{E}_3$ uniquely. We have, by this procedure, verified that the ansatz (4.13) leads to a unique solution from a known $\ddot{\sigma}^0$ and is thus consistent. An alternative point of view would be to consider D_3 and H_3 as freely specifiable and determine the other coefficients and $\ddot{\sigma}^0$ by (4.14), (4.15) and (4.16). It should be pointed out in this connection that it is H_3 (and E_3, G_3, F_3 and $\mathscr{E}_3$) which determines the curvature tensor. (See (7.13).)

Our ansatz for Z^0_{ab} and Z^0_{abc} is, of course, not the most general one; as we stated earlier, it amounts to the imposition of Riemann normal coordinates (to third order) about the origin. In the next Section, we shall use the above results to show that the norm (2.15) defines a Riemannian structure on $\mathcal{H}$-space, and to compute the Riemann tensor in terms of quantities on null infinity; the significance of the coordinate conditions will then be apparent.

5. The metric tensor

We next examine the metric on $\mathcal{H}$-space, given by the formula (2.15)

$$\frac{1}{\|u\|^2} = \frac{1}{8\pi}\oint \frac{\mathrm{d}S}{U^2}, \quad \mathrm{d}S = \frac{2}{\mathrm{i}}\frac{\mathrm{d}\zeta \wedge \mathrm{d}\tilde{\zeta}}{(1+\zeta\tilde{\zeta})^2}, \tag{5.1}$$

where $U(\zeta, \tilde{\zeta})$ defines a 'tangent vector' at Z in $\mathcal{H}$-space, being subject to

$$\eth^2 U = \dot{\sigma}^0 U. \tag{5.2}$$

We shall show that (5.1) is quadratic and is in fact, equivalent to the somewhat different definition given in (Penrose 1976) for the non-linear graviton construction. Then, having done so, we can appeal to the results of that construction. In particular, it follows without further calculation that $\mathcal{H}$ satisfies the Einstein vacuum equations (whether or not the physical space-time M does – provided only that the normal asymptotic conditions hold) and that the curvature is left-flat (or right-flat, in the case of $\tilde{\mathcal{H}}$). In the next two sections we give independent derivations of these curvature properties.

In fact, there are certain difficulties about (5.1). The integration is to be performed over a suitable contour in the $(\zeta, \tilde{\zeta})$-space. But there is no guarantee that such a contour will exist for a given U satisfying (5.2). This is because only a neighbourhood of the anti-holomorphic diagonal $\tilde{\zeta} = \bar{\zeta}$ can be used and the zeros of U must also be avoided. If U is 'close enough' to representing a real timelike vector at Z, then there is no problem because then the zeros of U lie far from the diagonal (figure 2*a*). The diagonal itself or any allowable deformation can be taken as the contour. If U is 'close' to representing a real spacelike vector, a contour exists (figure 2*b*). If U is null then we expect 'pinching' of the contour because $\|u\|^2 = 0$. Such pinching occurs in the allowable region of the $(\zeta, \tilde{\zeta})$-space only if U represents a null vector which is 'close enough' to a real null vector (figure 2*c*). But for a general complex null vector no point of 'pinching' actually occurs in this region (figure 2*d*). For any non-null U nearby, the position of the zero of U will not differ in any essential way from the situation in the null case of figure 2*d*. This illustrates the fact (which can be independently verified) that no allowable contour can exist for this situation *whether or not* the vector is null. Thus, definition (5.1) cannot be directly applied for *all* complex tangent vectors at a point of $\mathcal{H}$. Instead, we adopt (5.1) for those tangent vectors at Z which are close to being real and timelike (assuming that $\mathcal{H}$ itself is close enough to Minkowski space for these

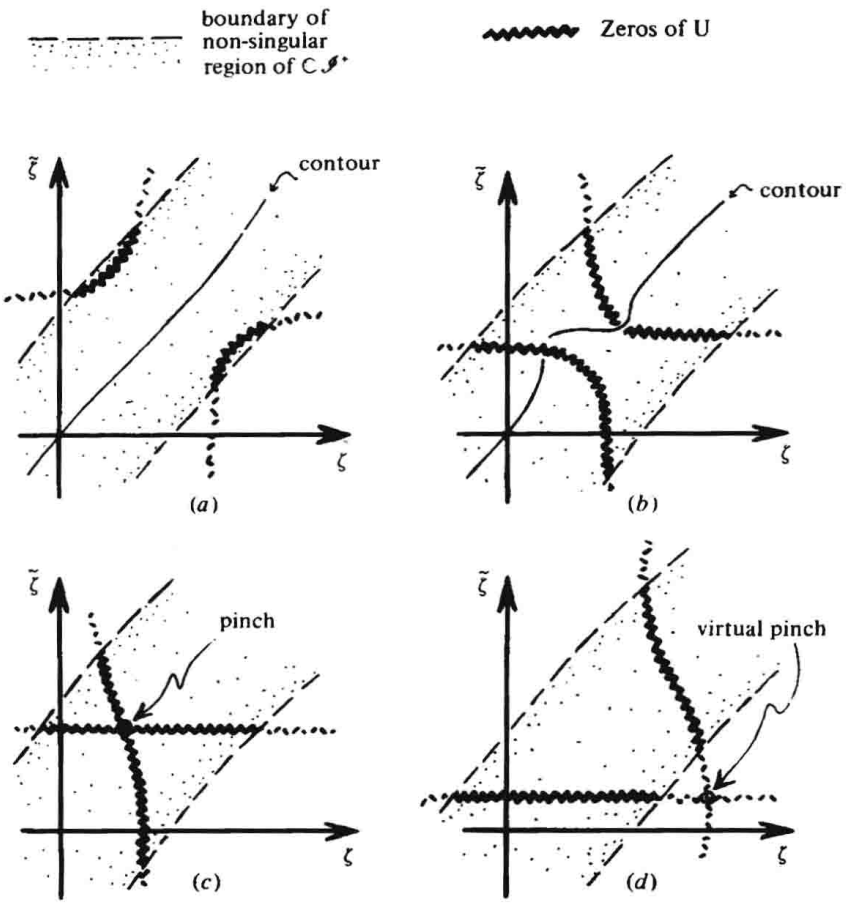


Figure 2. Plane view of a good cut. (*a*) U close to being real and timelike. (*b*) U close to being real and spacelike. (*c*) U null and close to being real. (*d*) U null and far from being real.

vectors to exist) and then extend the definition to all other tangent vectors at Z by analytic continuation.

The definition of the metric as given in (Penrose 1976) does not depend upon contour integration. In the present context, it is most easily stated in terms of the bracket

$$\{U, V\} = U\eth V - V\eth U. \tag{5.3}$$

This satisfies

$$\eth\{U, V\} = 0 \tag{5.4}$$

by virtue of

$$\eth^2 U = \dot{\sigma} U; \quad \eth^2 V = \dot{\sigma} V. \tag{5.5}$$

Now $\{U, V\}$ has spin weight one, so (5.4) implies that

$$\{U, V\} = \frac{(\tilde{\zeta}-\alpha)(\tilde{\zeta}-\beta)\gamma}{1+\zeta\tilde{\zeta}} \tag{5.6}$$

for some complex α, β, γ (the cases $\alpha = \infty$, $\beta = \infty$ being dealt with by a limiting argument). The expression

$$\frac{(\alpha-\beta)(1+\zeta\tilde{\zeta})}{(\tilde{\zeta}-\alpha)(\zeta-\beta)}\{U, V\} = \gamma(\alpha-\beta) \tag{5.7}$$

is thus independent of $\tilde{\zeta}$ as well as of ζ and, in the case of *null* vectors U, V *is* the (Penrose 1976) definition of the *scalar product* $g(U, V)$, from which the general metric definition follows by linearity. To avoid a circularity, we need an independent criterion for recognizing when U and V are to be interpreted as *null*. According to the twistor approach, two holomorphic curves in $P\mathcal{T}^*$ represent points in $\mathcal{H}$ which are null separated if the curves intersect. This implies that U represents a null vector if the two corresponding infinitesimally separated holomorphic curves in $P\mathcal{T}^*$ intersect. The intersection point defines a dual twistor line on the Z-cut which is a zero of U. The presence of a $\tilde{\zeta}$ = const. line of zeros in U is thus the required criterion for U to represent a null vector. Alternatively, the independent definition of nullness can be avoided by using the more general formula

$$((g(U, V))^2 - g(U, U)\, g(V, V))^{\frac{1}{2}} = \gamma(\alpha - \beta) \tag{5.8}$$

in place of (5.7), which then determines the metric up to sign.

To relate this definition (either (5.8) or (5.7)) of the metric, to (5.1) is, unfortunately, not quite direct. One method is to observe that

$$\frac{1}{U^2} = \eth\left(\frac{V}{U\{U, V\}}\right).$$

This provides an explicit partial integration of (5.1), to be performed as a definite integral in ζ leaving a contour integral in $\tilde{\zeta}$. Substituting the form (5.6) for $\{U, V\}$ then leads to an expression for $g(U, U)$. Interchanging U and V one calculates $g(V, V)$ and then $g(U+V, U+V)$ and it is then possible to obtain (5.8) showing that the two definitions agree. However, it is difficult to use this method for a rigorous demonstration when the allowable region in the $(\zeta, \tilde{\zeta})$-space may be narrow.

More satisfactory is to adopt a change of variable where $(\zeta, \tilde{\zeta})$ is replaced by $(\zeta^*, \tilde{\zeta})$, the new complex coordinate ζ^* referring to the *local* primed spin-space at the point Z in $\mathcal{H}$, i.e. if $\pi_{A'}$ is a primed spinor, $\zeta^* = \pi_{0'}/\pi_{1'}$. (The local unprimed spin-space agrees with the asymptotic unprimed spin-space, explaining why $\tilde{\zeta}(\tilde{\zeta} = \lambda_0/\lambda_1)$ can remain unchanged.) The primed spin-space at Z is however purely local and does not agree with the asymptotic spin-space. We can obtain a clearer understanding of this by considering the good cut representing the point Z in $\mathcal{H}$ and the neighbouring good cut representing a null tangent vector at Z (or null separated neighbouring point Z'). The two good cuts intersect along two curves, the straight line $\tilde{\zeta}$ = const. and a second curve. If now the tangent vector Z', represented by $Z^{AA'} = \lambda^A \pi^{A'}$ is varied, first by varying the $\pi^{A'}$ and then the λ^A, we obtain on the good cut Z (or $(\zeta, \tilde{\zeta})$ plane) two sets of intersection curves $\tilde{\zeta}$ = const. and $\zeta^* = \zeta^*(\zeta, \tilde{\zeta})$ = const. (See figure 3.)

We can now choose a new (S^2) coordinate ζ^* on the allowable region of Z which (provided the situation is close enough to the Minkowskian one) gives with $\tilde{\zeta}$, a new global holomorphic labelling of the allowed region with the ζ^* = const. lines representing the lines of constant primed projective spinors at $Z \in \mathcal{H}$. Thus $(\zeta^*, \tilde{\zeta})$

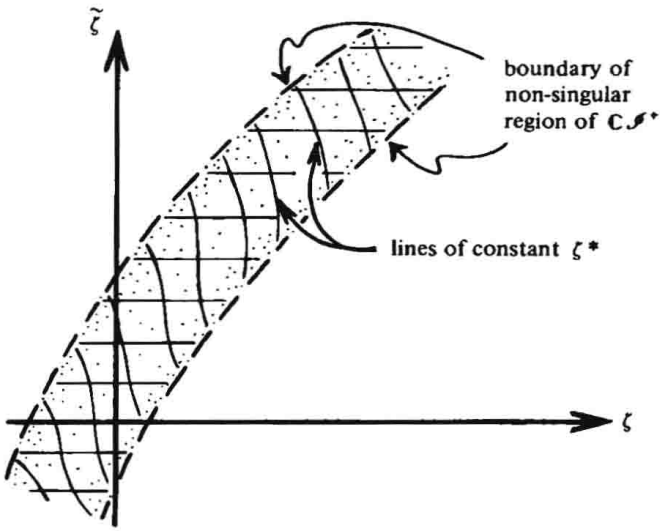


FIGURE 3. Plane view of a good cut showing the lines of constant ζ^*.

are the standard complex stereographic angular coordinates labelling the complex null directions at Z. We have $\zeta^* = \zeta^*(\zeta, \tilde{\zeta})$ single valued in the allowed region.

To find such a transformation, we consider a V_0 and W_0 from (4.5) with the V_0 being sufficiently close to real and timelike, so that it has no zeros in the allowed $(\zeta, \tilde{\zeta})$ region (cf. figure 2*a*). The transformation is then given by

$$\zeta^* = \frac{\zeta + W_0}{1 - W_0 \tilde{\zeta}}. \tag{5.9}$$

In the new coordinates $(\zeta^*, \tilde{\zeta})$ (4.6) becomes

$$Z^0_{\mathfrak{a}} V_0^{-1} = l^*_{\mathfrak{a}} = \frac{1}{\sqrt{2}}\left(1, \frac{-(\zeta^* + \tilde{\zeta})}{1 + \zeta^* \tilde{\zeta}}, \frac{-\mathrm{i}(\zeta^* - \tilde{\zeta})}{1 + \zeta^* \tilde{\zeta}}, \frac{1 - \zeta^* \tilde{\zeta}}{1 + \zeta^* \tilde{\zeta}}\right) \tag{5.10}$$

from which after a simple calculation we obtain

$$\frac{1}{8\pi} \oint \frac{\mathrm{d}S}{U^2} = \frac{1}{8\pi} \oint \frac{\mathrm{d}S^*}{U^{*2}} \tag{5.11}$$

with

$$\left.\begin{aligned} U &= u^{\mathfrak{a}} Z_{\mathfrak{a}}(\zeta, \tilde{\zeta}), \\ U^* &= u^{\mathfrak{a}} l^*_{\mathfrak{a}}(\zeta^*, \tilde{\zeta}), \\ \mathrm{d}S &= \frac{2}{\mathrm{i}} \frac{\mathrm{d}\zeta \wedge \mathrm{d}\tilde{\zeta}}{(1 + \zeta\tilde{\zeta})^2}, \\ \mathrm{d}S^* &= \frac{2}{\mathrm{i}} \frac{\mathrm{d}\zeta^* \wedge \mathrm{d}\tilde{\zeta}}{(1 + \zeta^*\tilde{\zeta})^2}, \end{aligned}\right\} \tag{5.12}$$

and

$$\{U, V\}^* \equiv U^* \eth^* V^* - V^* \eth^* U^* = \{U, V\} \tag{5.13}$$

with

$$\eth^* = V_0^2 \eth. \tag{5.14}$$

Note that now in the $(\zeta^*, \tilde{\zeta})$ system, the integral expression and the bracket expression are identical to those found in the Minkowski space case (see (2.2) and (2.5).) From (5.11) it immediately follows that the norm $\|u\|^2$ is quadratic and also after a brief calculation using (5.13) it is checked that the metric given by (5.8) is equal to that given by (5.1).

An important point should be noted here. If we construct, from the four independent solutions of (4.8), i.e. from the $l_{\mathbf{a}}$ of (4.9), the following sets of four functions

$$m_{\mathbf{a}} = \eth l_{\mathbf{a}},$$
$$\tilde{m}_{\mathbf{a}} = \tilde{\eth} l_{\mathbf{a}},$$
$$n_{\mathbf{a}} = l_{\mathbf{a}} + \eth\tilde{\eth} l_{\mathbf{a}}$$

one can easily show that the metric from (5.1) becomes

$$g_{\mathbf{ab}} = 2l_{(\mathbf{a}} n_{\mathbf{b})} - 2m_{(\mathbf{a}} \tilde{m}_{\mathbf{b})} = \eta_{\mathbf{ab}} \tag{5.15}$$

for any value of the ζ and $\tilde{\zeta}$. Though this result has been obtained in the Riemann normal coordinates it is easily seen to be true in any $\mathcal{H}$-space coordinate system. This constitutes the proof that the metric given here, either (5.1) or (5.8) is the same as that given in Ko *et al.* (1977). It also follows immediately from (4.6) and (4.7) with (5.15) that both $Z_{,\mathbf{a}}$ and $\eth Z_{,\mathbf{a}}$ for fixed but arbitrary ζ and $\tilde{\zeta}$ are null vector fields on $\mathcal{H}$-space.

6. The Riemann tensor: method I

We now describe a straightforward but lengthy method of obtaining the curvature properties of $\mathcal{H}$-space. In what follows, we regard Z at fixed ζ and $\tilde{\zeta}$ as a scalar function on $\mathcal{H}$-space and denote the nth $\mathcal{H}$-space covariant derivative of Z by $Z_{\mathbf{a}_1 \ldots \mathbf{a}_n}$ (in particular $Z_{\mathbf{a}} = Z_{,\mathbf{a}}$).

Using the fact that (5.1) defines a quadratic function of the tangent vector $u^{\mathbf{a}}$, we can obtain an expression for the metric components $g_{\mathbf{ab}}$ by differentiation with respect to $u^{\mathbf{a}}$ under the integral. Thus with

$$N = \left[\frac{1}{8\pi} g_{\mathbf{ab}} u^{\mathbf{a}} u^{\mathbf{b}}\right]^{-1} = \oint \mathrm{d}S (Z_{\mathbf{a}} u^{\mathbf{a}})^{-2} \tag{6.1}$$

we have

$$u_{\mathbf{a}} = g_{\mathbf{ab}} u^{\mathbf{b}} = 8\pi N^{-2} \oint \mathrm{d}S (Z_{\mathbf{a}} u^{\mathbf{b}})^{-3} Z_{\mathbf{a}} \tag{6.2}$$

$$g_{\mathbf{ab}} = 8\pi \left[4N^{-3} \left\{\oint \mathrm{d}S (Z_{\mathbf{c}} u^{\mathbf{c}})^{-3} Z_{\mathbf{a}}\right\} \left\{\oint \mathrm{d}S (Z_{\mathbf{d}} u^{\mathbf{d}})^{-3} Z_{\mathbf{b}}\right\} - 3N^{-2} \oint \mathrm{d}S (Z_{\mathbf{c}} u^{\mathbf{c}})^{-4} Z_{\mathbf{a}} Z_{\mathbf{b}}\right]. \tag{6.3}$$

The fact that the left-hand side of (6.3) is independent of $u^{\mathbf{a}}$ implies that we can substitute any tangent vector whatsoever into the integrals on the right without changing their values. It is convenient for subsequent calculations to use a vector field $v^{\mathbf{a}}$ such that, at the origin, $Z_{\mathbf{m}} v^{\mathbf{m}} = V_0$ and which satisfies

$$v^{\mathbf{a}}{}_{,\mathbf{b}} = 0 \tag{6.4}$$

in the coordinate system of § 4. The field v^a is thus fixed to third order in a Taylor series about the origin. We also define in the neighbourhood of the origin

$$V = Z_a v^a. \tag{6.5}$$

Equation (5.8) now takes the form

$$g_{ab} = 8\pi\left[4N^{-3}\left\{\oint \mathrm{d}S V^{-3} Z_a\right\}\left\{\oint \mathrm{d}S V^{-3} Z_b\right\} - 3N^{-2}\oint \mathrm{d}S V^{-4} Z_a Z_b\right]. \tag{6.6}$$

The strategy is to differentiate (6.6) with respect to the coordinates to obtain $g_{ab,c}$ and $g_{ab,cd}$ and to evaluate these expressions at the origin using the Taylor series solution for the good-cut function from § 4. The resulting expressions clearly yield integral formulae for the connection coefficients Γ^a_{bc} and the components R_{abcd} of the curvature tensor at the origin, in terms of derivatives of the good-cut function. As the calculation of these quantities is long and cumbersome, while relatively straightforward, we shall only sketch the method.

The first differentiation of (6.6) yields

$$\begin{aligned}
\frac{1}{8\pi} g_{ab,c} = {} & 6N^{-4} N_a N_b \oint \mathrm{d}S V^{-3} Z_{cm} v^m \\
& - 2N^{-3}\left[N_a \oint \mathrm{d}S V^{-3} Z_{bc} + N_b \oint \mathrm{d}S V^{-3} Z_{ac}\right] \\
& + 6N^{-3}\left[N_a \oint \mathrm{d}S V^{-4} Z_b Z_{cm} v^m + N_b \oint \mathrm{d}S V^{-4} Z_a Z_{cm} v^m\right] \\
& - 12N^{-3}\left[\oint \mathrm{d}S V^{-4} Z_a Z_b\right]\left[\oint \mathrm{d}S V^{-3} Z_{cm} v^m\right] \\
& - 3N^{-2}\left[\oint \mathrm{d}S V^{-4}(Z_a Z_{bc} + Z_b Z_{ac})\right] \\
& - 12N^{-2}\left[\oint \mathrm{d}S V^{-5} Z_a Z_b Z_{cm} v^m\right]
\end{aligned} \tag{6.7}$$

with

$$N_a = N_{,a} = -2\int \mathrm{d}S V^{-3} Z_a.$$

By substituting (4.10) into (6.7) and using (4.11) repeatedly when integrating by parts, one can after a lengthy calculation, show that

$$g_{ab,c} = 0 \tag{6.8}$$

and hence

$$\Gamma^a_{bc} = 0 \tag{6.9}$$

at the origin. This substantiates our claim that the coordinate conditions imposed in § 4 yield Riemann normal coordinates to second order.

The calculation of the Riemann tensor at the origin is again straightforward but very lengthy: (6.7) is differentiated, yielding a very long integral expression for $g_{ab,cd}$, from which the curvature tensor can be evaluated in the form

$$R_{abcd} = \tfrac{1}{2}(g_{ad,bc} + g_{bc,ad} - g_{ac,bd} - g_{bd,ac}). \tag{6.10}$$

The connection terms are absent because of (6.9). We thus have an expression for the curvature tensor involving integrals over Z_a, Z_{ab}, and Z_{abc}. By substituting (4.6), (4.10) and (4.13) and integrating by parts many times using (4.5), (4.11), (4.14) and (4.15), one obtains the following expression

$$\begin{aligned} R_{abcd} = -\frac{1}{\pi}\oint \mathrm{d}S\{&(V_0^4\ddot{\sigma}^0 + 3V_0^3 A_2\dot{\sigma}^0)\, l^*_{[a}\tilde{m}^*_{b]}\, l^*_{[c}\tilde{m}^*_{d]} \\ &+ V_0^3 B_2\ddot{\sigma}^0[l^*_{[a}\tilde{m}^*_{b]}(l^*_{[c}n_{d]} + m_{[c}\tilde{m}^*_{d]}) + (l^*_{[a}n_{b]} + m_{[a}\tilde{m}^*_{b]})\, l^*_{[a}\tilde{m}^*_{d]}] \\ &+ \tfrac{1}{6}V_0^3 C_2\ddot{\sigma}^0[l^*_{[a}\tilde{m}^*_{b]}m_{[c}n_{d]} + (l^*_{[a}n_{b]} + m_{[a}\tilde{m}^*_{b]})(l^*_{[c}n_{d]} + m_{[c}\tilde{m}^*_{d]}) + m_{[a}n_{b]}\, l^*_{[c}\tilde{m}^*_{d]}]\}. \end{aligned} \tag{6.11}$$

From (5.15), (4.12) and (4.7), it is easily seen that l^*_a, m_a, n_a and $\tilde{m}^*_a$ form a null tetrad system for each fixed value of ζ and $\tilde{\zeta}$. From this it follows simply that the curvature tensor is trace free, i.e.

$$R_{ab} = 0 \tag{6.12}$$

and further that it is self-dual, i.e.

$$^*R_{abcd} \equiv \tfrac{1}{2}e_{ab}{}^{mn}R_{mncd} = iR_{abcd}. \tag{6.13}$$

These are exactly the properties which can be obtained from the asymptotic twistor construction discussed in § 3. Here, however, we obtain the explicit dependence of the curvature on the asymptotic radiation field. It is easily verified that R_{abcd} vanishes in the (rather uninteresting) case of constant radiation fields, i.e. $\dot{\sigma}^0 \neq 0$, $\ddot{\sigma}^0 = \dddot{\sigma}^0 = 0$. Thus the Riemann tensor (or Weyl tensor) appears to depend essentially on the rate of change of the radiation. In a very rough sense, the curvature may be thought of as the quadrupole part of $\ddot{\sigma}^0$, this becoming precisely true if $\dot{\sigma}^0$ and $\ddot{\sigma}^0$ vanish at the point under consideration.

We remarked in § 2 that we might equally well have considered $\tilde{\mathcal{H}}$-spaces. In that case, the curvature tensor is again Ricci-flat but anti-self-dual. Thus $\mathcal{H}$-spaces (sometimes called left-flat $\mathcal{H}$-spaces) have right-handed or self dual curvatures and the reverse for $\tilde{\mathcal{H}}$-spaces.

7. The Riemann tensor: method II

An alternative method of deriving the curvature tensor and studying its properties is obtained from a formalism based on the bracket (5.3). Though the method is not as obvious and direct as the method of § 6, the details of the calculations are dramatically simplified. We only sketch the method.

The bracket may be defined as follows: Let $T(z^a, \zeta, \tilde{\zeta})$ and $U(z^a, \zeta, \tilde{\zeta})$ be solutions of $\eth^2 Z_a = \dot{\sigma}^0 Z_a$ at the point z^a, i.e. they are of the form

$$T = Z_a t^a;\ U = Z_a u^a \tag{7.1}$$

then the bracket is (as in § 3)

$$\{T, U\} \equiv T\eth U - U\eth T. \tag{7.2}$$

By using (4.6) and (4.7) we easily obtain

$$\{T, U\} = t^a u^b \{Z_a, Z_b\}; \tag{7.3}$$

where

$$\{Z_a, Z_b\} = 2l_{[a} m_{b]}. \tag{7.4}$$

With

$$\tilde{m}_a = \eth l_a;\ n_a = l_a + \eth\tilde{\eth} l_a$$

we have by direct calculation that

$$\tilde{\eth}\{Z_a, Z_b\} = 2(l_{[a} n_{b]} - m_{[a} \tilde{m}_{b]}) \tag{7.5}$$

and

$$\tilde{\eth}^2\{Z_a Z_b\} = 4\tilde{m}_{[a} n_{b]}. \tag{7.6}$$

In § 5 we noted, for each value of ζ and $\tilde{\zeta}$, that l_a, $\tilde{m}_a$, m_a, n_a define (at each z^a) a null tetrad system and hence that

$$g_{ab} = 2l_{(a} n_{b)} - 2m_{(a} \tilde{m}_{b)}. \tag{7.7}$$

In addition it is easily seen that

$$e_{abcd} = 24\mathrm{i} l_{[a} m_b \tilde{m}_c n_{d]}. \tag{7.8}$$

Putting (7.7) and (7.8) together, we have

$$\begin{aligned} 2g_{a[c} g_{d]b} - \mathrm{i} e_{abcd} = {} & 8l_{[a} m_{b]} \tilde{m}_{[c} n_{d]} \\ & - 4(l_{[a} n_{b]} - m_{[a} \tilde{m}_{b]})(l_{[c} n_{d]} - m_{[c} \tilde{m}_{d]}) \\ & + 8\tilde{m}_{[a} n_{b]} l_{[c} m_{d]}. \end{aligned} \tag{7.9}$$

Finally substituting (7.4), (7.5) and (7.6) into (7.9) we have

$$\begin{aligned} 2g_{a[c} g_{d]b} - \mathrm{i} e_{abcd} = {} & \{Z_a, Z_b\} \tilde{\eth}^2\{Z_c, Z_d\} \\ & - \tilde{\eth}\{Z_a, Z_b\} \tilde{\eth}\{Z_c, Z_d\} + \tilde{\eth}^2\{Z_a, Z_b\}\{Z_c, Z_d\}. \end{aligned} \tag{7.10}$$

It can be shown that (the Z_a being known) (7.10) determines the metric up to sign. Furthermore the requirement that

$$\nabla_a g_{bc} = 0;\ \nabla_a e_{bcde} = 0, \tag{7.11}$$

implies that

$$\nabla_a \{Z_b, Z_c\} = 0 \tag{7.12}$$

from which the Christoffel symbols can be directly calculated in terms Z_a and Z_{ab}. We emphasize that though the final results are the same as in § 6, (in fact if the Z_a and Z_{ab} of § 4 are used here we obtain $\Gamma^a_{bc} = 0$) the expressions for the connection is algebraic in Z_a and Z_{ab} and does not involve integrals.

In addition, since $\{Z_a, Z_b\}$ for fixed ζ and $\tilde{\zeta}$ is an anti-self-dual bivector field and since also at a point all anti-self-dual bivectors (up to proportionality) are obtained from $\{Z_a, Z_b\}$ by varying ζ and $\tilde{\zeta}$, one has immediately from (7.12) that there is a global parallelism of anti-self-dual bivectors. The integrability conditions for this imply immediately that the Ricci tensor vanishes and the curvature tensor is left-conformally flat, a result in agreement with that of the previous section.

As a final remark we point out that since the connection can be expressed as an algebraic function of Z_a and Z_{ab}, so can the curvature tensor be expressed as a function of Z_a, Z_{ab} and Z_{abc}. If the results of § 4 for Z_a, Z_{ab} and Z_{abc} are used here one obtains

$$\begin{aligned} -\tfrac{1}{2}R_{abcd} = &-\mathscr{E}_3\tilde{L}_{ab}\tilde{L}_{cd} + E_3[\tilde{L}_{ab}\tilde{M}_{cd} + \tilde{L}_{cd}\tilde{M}_{ab}] \\ &+ F_3[\tilde{L}_{ab}\tilde{N}_{cd} + \tilde{M}_{ab}\tilde{M}_{cd} + \tilde{N}_{ab}\tilde{L}_{cd}] \\ &+ G_3[\tilde{M}_{ab}\tilde{N}_{cd} + \tilde{N}_{ab}\tilde{M}_{cd}] \\ &+ H_3 N_{ab} N_{cd} \end{aligned} \tag{7.13}$$

with $\tilde{L}_{ab} = l^*_{[a}\tilde{m}^*_{b]}$, $\tilde{M}_{ab} = l^*_{[a}n_{b]} + m_{[a}\tilde{m}^*_{b]}$, $\tilde{N}_{ab} = m_{[a}n_{b]}$.

In order to obtain agreement with (6.11), this expression must be integrated over the sphere. The potentials $\mathscr{E}_3, E_3 \ldots H_3$ are to be eliminated by integration by parts using (4.14) and (4.15).

It should be pointed out that, recently, additional methods for studying the curvature properties of $\mathscr{H}$-space have been developed that do not use integral expressions for the metric and curvature. (Ko *et al.* 1977; Hansen & Ludvigsen 1977).

References

Bondi, H., van der Burg, M. G. J. & Metzner, A. W. K. 1962 *Proc. R. Soc. Lond.* A **269**, 21.

Brieskorn, E. 1964 *Math. Annln* **155**, 184.

Gunning, R. C. 1966 *Lectures on Riemann surfaces*. Princeton U.P.

Hansen, R. O. & Ludvigsen, M. 1977 *Gen. Rel. Grav.* **8**, 761.

Hansen, R. O. & Newman, E. T. 1975 *Gen. Rel. Grav.* **6**, 361.

Ko, M., Newman, E. T. & Tod, K. P. 1977 *Asymptotic structure of space–time* (eds Esposito & Witten). New York: Plenum Press.

Kodaira, K. 1962 *Ann. Math.* **75**, 146.

Kodaira, K. 1963 *Am. J. Math.* **85**, 79.

Morrow, J. & Kodaira, K. 1971 *Complex manifolds*, Holt, Rinehart and Winston.

Newman, E. T. 1976 *Gen. Rel. Grav.* **7**, 107.

Newman, E. T. & Penrose, R. 1966 *J. math. Phys.* **7**, 863.

Penrose, R. 1975 In *Quantum gravity: an Oxford Symposium* (eds Isham, Penrose & Sciama). Oxford University Press.

Penrose, R. 1976 *Gen. Rel. Grav.* **7**, 31.

Penrose, R. & MacCallum, M. A. H. 1973 *Phys. Rep.* **6** C, 242.

Penrose, R. 1974 *Group theory in non-linear problems* (ed. A. O. Barut), Dordrecht, Boston: Reidel.

Sachs, R. K. 1962 *Phys. Rev.* **128**, 2851.

Wells, R. O. 1973 *Differential analysis on complex manifolds*. Englewood Cliffs: Prentice Hall.

98
THE GEOMETRY OF THE UNIVERSE

Mathematics today, twelve informal essays. Ed. L. A. Steen, Springer, New York, 1978

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

The Geometry of the Universe

Roger Penrose

One of the most fruitful sources of mathematical intuition is physical space. For not only does physical space provide us with the basic concepts of Euclidean geometry, but it also gives us a pictorial framework for visualizing the very much more general types of space that occur continually throughout mathematics. Moreover, it was the picture of physical space that led to those key ideas of mathematical analysis: continuity and smoothness. Indeed, even the very basic mathematical notion of real number originated from measurement of spatial separation—and of time intervals too, these being, as Albert Einstein's relativity has told us, geometrical quantities again, whose measurement is essentially bound up with that of space. So it comes as a shock when we also learn from relativity that our now cherished notion of Euclidean geometry does not, after all, describe physical space in the most accurate way. Yet, from these Euclidean beginnings, a more subtle and flexible geometry, known as differential geometry, has grown to maturity. It is in terms of this geometry that Einstein's theory finds expression. And now, more than sixty years after general relativity was first put forward as a daring original view of the world, the theory stands in excellent agreement with observation. So if we wish to understand how the world is shaped, we must come to terms with this theory.

In this article I shall first introduce the simplest models of non-Euclidean geometry and then indicate how the powerful general machinery (manifolds, tensors, tangent spaces, geodesics) of differential geometry may be constructed. In so doing, I shall use some of the more modern ideas for this that were not available to Einstein in his day. Finally, I shall outline the basic physical notions of both special and general relativity, describing the geometric content of Einstein's equations and taking the discussion as far as black holes and cosmology.

The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

The physical theory of general relativity could not have evolved were it not for the work of many generations of mathematicians (in particular, Carl Friedrich Gauss and Bernhard Riemann) who were able to free geometry from its earlier imprisonment in Euclidean rigidity. But this debt to pure geometry that relativity had owed has now been amply repaid. For many of the ideas of the modern subject of differential geometry received their initial stimulus from concepts arising from Einstein's general relativity.

Thus we have, in the physics of relativity and in the mathematics of differential geometry, a supreme example which illustrates how physics and mathematics can enrich one another. Here we can begin to see something of that profound interplay between the workings of the natural world and the laws and sensitivity of thought—an interplay which, as knowledge and understanding increase, must surely ultimately reveal a yet deeper interdependence of the one upon the other.

Non-Euclidean Geometries

But what is differential geometry? I shall begin by describing what was historically the first of the non-Euclidean geometries—Lobachevskian (or hyperbolic) geometry—which deviates the least from that of Euclid. Just one of the Euclidean postulates is abandoned, namely the fifth: "If a straight line falling on two straight lines make the interior angles on the same side less than two right angles, then the two straight lines if produced indefinitely meet on that side on which the angles are less than two right angles." Lobachevskian geometry was, in effect, explored in considerable detail by the Jesuit Geralamo Saccheri in a work published in 1733, though his aim had really been to show that Euclid's fifth postulate was consequent upon the others. But it was the great German mathematician Carl Friedrich Gauss (1777–1855) who first clearly realized that in fact a consistent alternative geometry to that of Euclid could be obtained in which the fifth postulate of Euclid was simply denied. Gauss did not publish his findings, and his main results in this area were independently rediscovered first (apparently) by his contemporary Ferdinand Schweikart, then by the Russian Nikolai Ivanovich Lobachevski (1793–1856) and then again by the Hungarian János Bólyai (1802–1860). As is not infrequently the case in mathematics, the name which becomes most commonly attached to an idea need not belong to the person who first thought of it. (But I wish to emphasize that, contrary to the sentiments apparently expressed in Tom Lehrer's witty satirical song, there is absolutely no suggestion that Lobachevski had heard of these ideas beforehand.)

Consider the two-dimensional case first. On way to proceed is to model the Lobachevskian plane in terms of ordinary Euclidean concepts but where the basic notions of "distance" and "straight" are interpreted in a new way. We shall name this model L^2, where the superscript "2" (and "1" in "S^1",

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

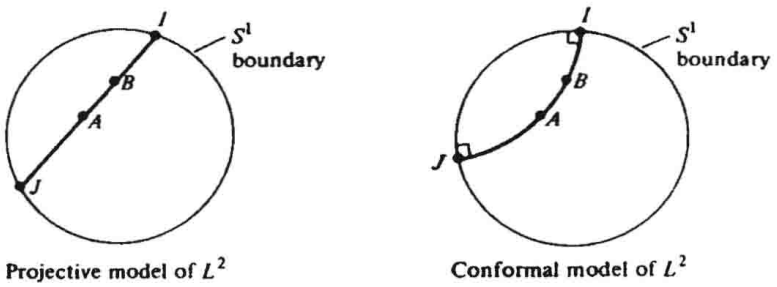


Figure 1. The Lobachevski Space L^2. "Lobachevskian distance AB" $= \epsilon \log_e (IA \cdot JB / IB \cdot JA)$ where IA, JB etc. denote *Euclidean* distances, and where $\epsilon = \frac{1}{2}$ for the projective model and $\epsilon = 1$ for the conformal model.

etc.) refers to the dimension of the space. (Read "L^2" as "L-two", not "L squared.") There are two standard representations of L^2 in the Euclidean plane E^2, the projective model and the conformal model. In each case, L^2 is represented as a disc in E^2—the interior of a circle S^1 of unit radius (see Figure 1). According to the projective model a "straight line" of Lobachevskian geometry is depicted as a chord, i.e., as a line segment drawn across the disc which is *straight* also with respect to the Euclidean geometry of E^2. According to the conformal model, a Lobachevskian "straight line" is depicted as an arc drawn across the disc which is *circular* with respect to the geometry of E^2 and which meets S^1 at right-angles. The Lobachevskian "distance" between two points A, B of the disc is given by a simple formula (see Figure 1) which, remarkably, has almost precisely the same form in each model.

While the projective model has the obvious advantage that the Lobachevskian "straightness" is correctly depicted, the conformal model has the slightly more subtle and perhaps more significant advantage that Lobachevskian "angles" are correctly depicted. This has the effect that small shapes are represented without distortion, only an appropriate magnification having to be applied at each point. Furthermore, Lobachevskian "circles" are always correctly depicted as circles in this model. Some very elegant woodcuts by the Dutch artist Mauritus C. Escher illustrate these facts in a striking way (see Figure 2), being accurate representations of the conformal model of L^2. (The idea of making such illustrations was suggested to Escher by the Canadian geometer H.S.M. Coxeter.) Notice that though the figures become very small and crowded towards the edge, they remain roughly true in their shape. The smaller the portion of each figure that is examined, the truer its shape remains. For that is what is meant by a conformal representation: arbitrarily small figures are accurately represented, with only a change in scale.

The Lobachevski world is an infinite world. For example, each devil in Escher's second woodcut is of equal Lobachevskian area, and there are infinitely many devils. The boundary circle S^1 represents infinity for L^2.

The conformal and projective models can be directly related to one

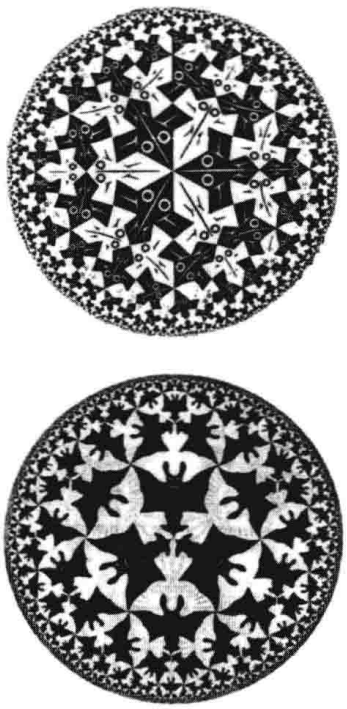
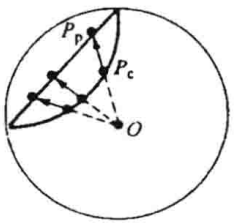


Figure 2. Escher's view of L^2.

another. If the two diagrams are superimposed, then the point P_p of the projective model is obtained from the corresponding point P_c of the conformal model by moving it radially outwards from the centre O by the amount shown in Figure 3. The fact that we have two different ways of representing L^2 in the Euclidean plane (there are in principle many more) illustrates an important point. It is the internal (or *intrinsic*) geometry of L^2 that really concerns us. Many different ways of representing this same geometry are possible, but the differences have to do with the representation only.

For another type of two-dimensional geometry, distinct from that of both L^2 and E^2, consider an ordinary spherical surface S^2 in three-dimensional Euclidean space E^3. Again Escher has depicted his vision of life in such a world in terms of interlocking angels and devils, this time in the form of a wooden model (see Figure 4). For a specific mathematical description we can take the sphere S^2 to have unit radius. The "straight lines" with respect to the intrinsic geometry are the great circles (i.e., circles on S^2 whose centers coincide with that of S^2). The measure of "distance" on S^2 is simply arc length along a great circle. The geometry of S^2 differs from Euclidean geometry a little more seriously than does that of L^2. For the "straight" lines of S^2 are closed up on themselves (they have, in mathematical terminology, the same *topology* as circles) and intersect one another in *pairs* of points rather than in single points. This last difference can be removed if we pass to the so-called elliptic space P^2, which is derived from S^2 by the identification of antipodal points. Each point of P^2, therefore, corresponds to *two* points of S^2, which lie

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978



$$\overrightarrow{OP}_p = \left(\frac{2}{1 + OP_c^2}\right) \overrightarrow{OP}_c \ .$$

Figure 3. The relation between the two L^2 models.

on a straight line through its center. We may think of each point of P^2 to be defined abstractly as such a pair of antipodal points of S^2. (Such abstract representations of spaces are common in mathematics. What is defined as a "point" in one space may just be some more elaborate structure in another. We shall see more examples of this later.)

If we try to represent the points of S^2 or P^2 globally on the Euclidean plane we have a difficulty. This is because the topology of S^2 and P^2 is different from that of E^2 or of any subset of E^2. But we shall see in a moment how to obtain a valid representation by adjoining some "points at infinity" to E^2. The topology of P^2 can be described as a "sphere with a cross-cap," as depicted in Figure 5. In Figure 6 we see how to get from the first picture of Figure 5 to the second. Since each point of P^2 appears twice in S^2, we consider, say, just the southern hemisphere of S^2, but where antipodal points on the equator must be identified.

Figure 4. Escher's model of an S^2 world.

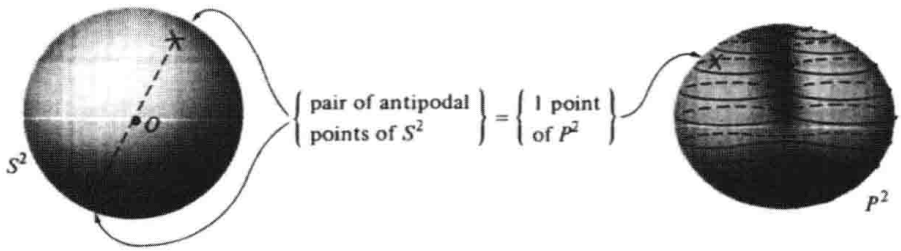


Figure 5. Spherical and elliptic geometry.

A standard conformal representation of S^2 on the plane is given by *stereographic projection*. Regard E^2 as the equatorial plane of S^2 and let N be the north pole. A point X on the sphere is projected from N to the point Z on E^2 (so N, X, and Z are all in line) as shown in Figure 7. Notice that as X approaches N, the line becomes a tangent of S^2 at N and is therefore parallel to E^2. The point Z has gone to "infinity" on E^2. Thus we have realized S^2 not as a subset of E^2 but as the whole of E^2 together with one extra point, adjoined at infinity, to represent N.

This stereographic correspondence is an important one. It is conformal in the same sense as was the conformal model of L^2. An angle between curves on S^2 projects to an equal angle between the corresponding projected curves; a small shape on S^2 projects to a similar small shape on E^2 (perhaps expanded or contracted). Furthermore (over and above conformality) any circle of any size on S^2 will project to an exact circle on E^2—except that if the circle on S^2 happens to pass through N, the projection is a straight line in E^2 (see Figure 8). A great circle on S^2 (i.e., a "straight line" for its intrinsic geometry) is represented on E^2 as a circle (or straight line) whose two intersections with the equator are diametrically opposite.

Let us now imagine N and the plane E^2 to be held fixed, but the sphere S^2 to be moved upwards until its center is the point N, so that it now touches E^2 at the point O. The projection, from N, of S^2 to E^2 now carries a *pair* of antipodal points X, X' on S^2 to a single point Z of E^2, except that the points of the entire equator are now carried to infinity (see Figure 9). Great circles on S^2 are now carried to straight lines on E^2. The model is projective, rather than conformal and represents P^2, rather than S^2 (since X and X' give the same point). Now E^2 is to have the points of an additional *line* adjoined at infinity (representing the equator of S^2).

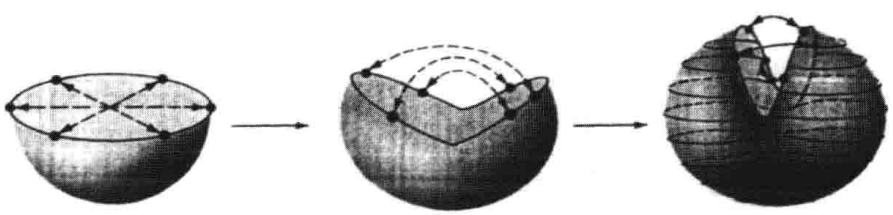


Figure 6. How the "cross-cap" topology of P^2 arises.

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

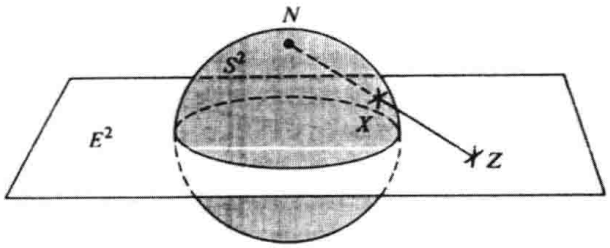


Figure 7. Stereographic projection of S^2.

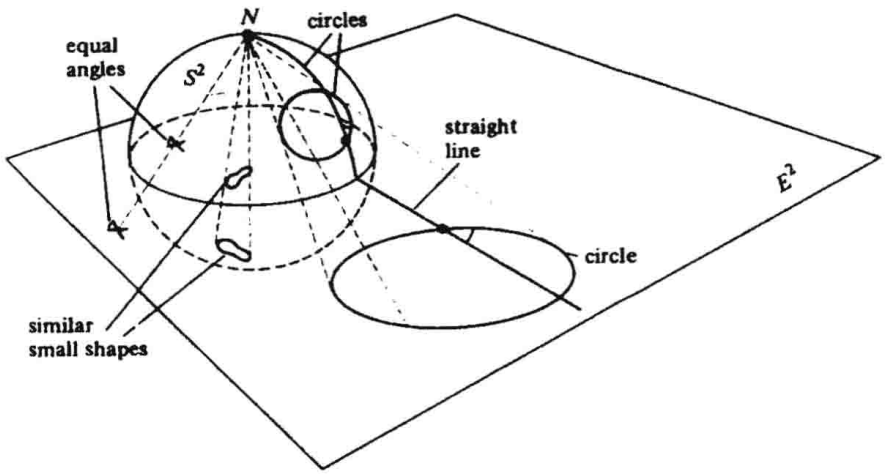


Figure 8. Effects of the stereographic projection.

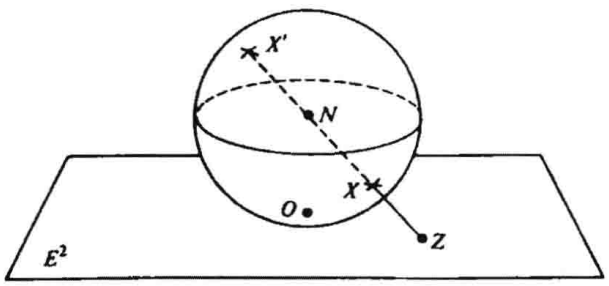


Figure 9. Projection of P^2 to E^2.

As in the case of our two models for L^2, these representations of S^2 and of P^2 can be related to one another by simple geometry (see Figure 10). Note that the formula relating the two models differs by only a sign change from the Lobachevskian one of Figure 3.

We have been considering particular ways of representing the geometries L^2, S^2, and P^2 in terms of a Euclidean description. But each of these geometries can be developed in its own way in an entirely "synthetic" manner, as

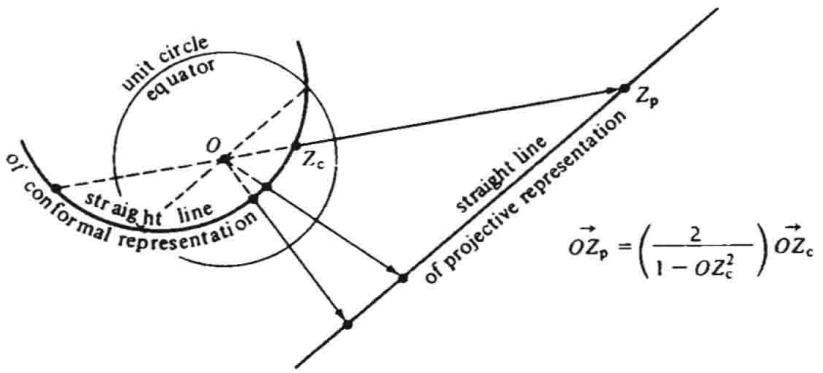


Figure 10. The relation between the S^2 and the P^2 models.

an alternative to the geometry of Euclid. Each has its own body of theorems, sometimes agreeing with those of Euclid, sometimes more complicated, occasionally simpler. A striking example in this last category is the formula for the area of a triangle discovered by Johann Heinrich Lambert (1728–1777) even before non-Euclidean geometry was established! If the angles are α, β, and γ (measured in radians, where 180° equals π radians) and the sides are "straight" with respect to the relevant geometry, then the area Δ is given by how much the sum of the angles differs from π, i.e., by

$$\Delta = \begin{cases} \pi - (\alpha + \beta + \gamma) & \text{for } L^2, \\ (\alpha + \beta + \gamma) - \pi & \text{for } S^2 \text{ or } P^2. \end{cases}$$

No such simple formula exists in the Euclidean case, for there we always have $\alpha + \beta + \gamma = \pi$ (= 180°) and a side length must be specified also in order that the area can be determined. This brings out another feature of the geometries L^2, S^2, and P^2, which is distinct from the Euclidean case, namely that there is now an absolute scale of length defined. The Euclidean property of similarity does not now hold. A large triangle is now not really like a small one, since the difference between π and the sum of its angles is a measure of its size. In fact, in the preceding formulas, units of length have been tacitly assumed to concur with this absolute scale. Thus S^2 was taken as a *unit* sphere, its total area being 4π; similarly, the area of P^2 is 2π. The scale of L^2 shows up in the fact that π is the upper bound of all possible areas of triangles. By recalibrating the unit of length we obtain Lobachevskian, spherical, or elliptic geometries whose quantitative departure from that of Euclid is as large or as small as we please.

These geometries are of especial interest not only because they represent the simplest deviations from the Euclidean case but also because they describe, according to mainstream modern cosmology, the best candidates for the large-scale spatial structure of the actual universe. But since physical space is three-dimensional rather than two-dimensional, we must consider the extensions to three dimensions of the models just introduced. In fact, this extension is quite straightforward. We can model the three-dimensional

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

Lobachevskian space L^3, again either by a projective or conformal representation, as the interior of a sphere S^2 in ordinary Euclidean three-space E^3, the Lobachevskian definition of "straightness" and of "distance" being the same as in the two-dimensional case. Thus, for example, in the conformal representation, a Lobachevskian "straight line" is depicted as a circular arc, in the 3-space inside S^2, which is perpendicular to S^2 at the two points I, J where it meets it; and the formula for the Lobachevskian distance between points A, B on the arc is as given in Figure 1. In fact it is the Lobachevski geometry L^3 that is most favoured by present (albeit somewhat indirect) observations. So why is it that we are not directly aware that the geometry of the spatial world probably accords more with that of Lobachevski than with that of Euclid? The reason lies in the fact that the magnitude of the absolute unit of length (the "radius of curvature of the spatial universe") is of the order of 10,000,000,000 light years which is approximately

60,000,000,000,000,000,000,000 miles!

Compared with this, the measurement of any terrestrial distance must be regarded as essentially infinitesimal. And the *infinitesimal* geometry of L^3 is Euclidean.

The S^3 and P^3 models must also be considered seriously in this connection. Cosmological observations do not yet have a good record of reliability; and the balance of informed opinion could easily shift from its present preference, namely L^3, to S^3 (or P^3). Again we can represent these geometries in terms of E^3, using a conformal or a projective description. The discussion is just as before—though it is harder actually to visualize the projections which achieve the representation, since this would require *four* Euclidean dimensions!

Two new features which do arise are, perhaps, worthy of note here. The first concerns the *orientability* of P^3, as opposed to the nonorientability of P^2. The question is whether a consistent sense of local "handedness" can be defined over the whole space. Consider S^2 first. At each point we can define a handedness (or orientation) in terms of a rotation sense in the surface about that point—say clockwise as we view the sphere from the outside (see Figure 11). But the antipodal correspondence is incompatible with preserv-

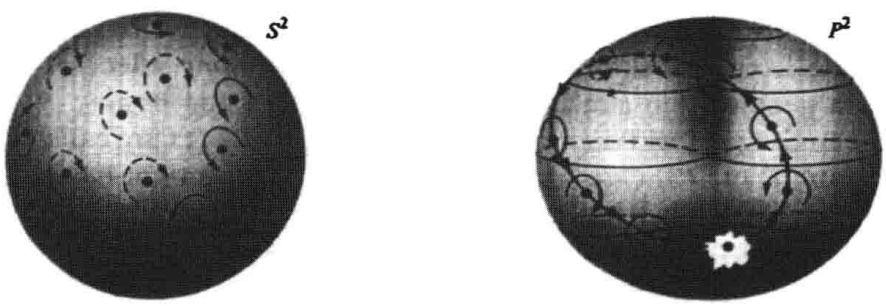


Figure 11. Orientability of S^2 and nonorientability of P^2.

ing this sense of orientation, so that when we identify antipodal points to produce P^2 we get a nonorientable surface (Figure 11) which cannot have a consistent continuous orientation assigned to it: by travelling around certain closed paths on P^2 we can return to our starting point but with the sense of handedness reversed.

The new feature that arises with P^3, however, is that this is *not* the case in three (or any odd number of) dimensions; that is, the antipodal correspondence on S^3 is here compatible with the orientation, so that P^3 is orientable. Thus, a space-traveller who travels from his home planet, in a P^3 world, and finally returns, will never find (as happened to the school-teacher in H.G. Wells's "The Plattner Story") that his sense of handedness is reversed in relation to the standard defined on his home planet. There are, however, other (nonhomogenous) three-dimensional models which are not orientable, so that if the space traveller in such a world left his right glove behind on the planet and took his left glove with him, he could find, on his return, that he had two identically rather than oppositely shaped gloves!

The second new feature arising is a more detailed one, which may seem (initially) like a curiosity. I mention it because it is actually important in a number of different contexts and is a good example of a recurring theme in differential geometry, as we shall see. Recall that the geometries L^2 and P^2 (or S^2) differ from that of Euclid with regard to the concept of parallel lines. Thus in the case of L^2, two straight lines may appear to start out parallel to one another, but as we move out along the lines, they begin to diverge away from one another more and more, farther and farther along the lines. On the other hand, with S^2 or P^2 the situation is opposite: two lines which start out seeming to be parallel will begin to converge towards one another and finally meet. If we pass to the three-dimensional geometry L^3, or to S^3 and P^3, this situation is exactly corresponding; but for S^3 (or P^3) a slightly different concept of "parallel" can be introduced for which the lines do *not* get closer together!

First consider a straight line l in ordinary Euclidean space E^3. Imagine l to be displaced parallel to itself by moving it out in a direction $\vec{d}$, perpendicular to itself. Now rotate the line slightly about the direction $\vec{d}$ as we move it out. The result is a line m which is *skew* to l. The distance between l and m will increase if we move outwards along them in either direction. Now apply the same procedure in the space S^3 (or P^3). By carefully adjusting the rate at which we rotate as we move m away from l, we can balance this *diverging* effect arising from the skewness against the *converging* effect of the non-Euclidean geometry in such a way that the lines l and m now remain the *same* distance apart all the way along. Such pairs of lines are called *Clifford parallels* on S^3 (or P^3) (see Figure 12) and can be of a right-handed or left-handed sort. (William Kingdon Clifford (1845–1879), an English geometer, had forcefully suggested that matter curves space!)

More remarkably, it is possible to fill up the whole of S^3 (or P^3) with a system of lines (great circles on S^3), all of which are "parallel" to one another in

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

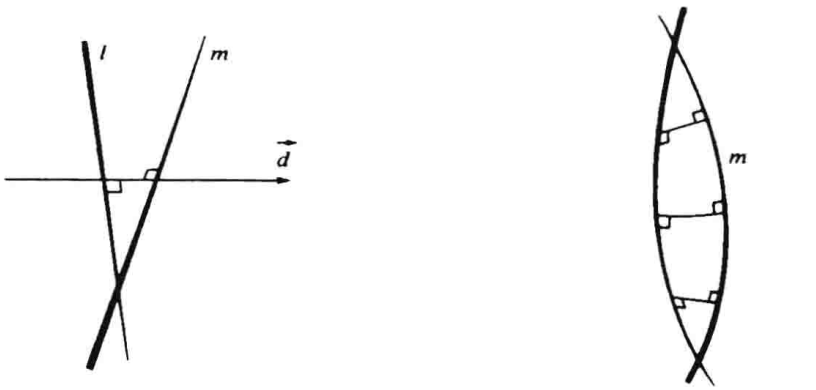


Figure 12. Clifford parallels.

this sense, where there is precisely one through each point of the space and where every pair is linked. Figure 13 illustrates such a complete system of Clifford parallels. The S^3 is represented in ordinary Euclidean space E^3 (together with one point at infinity) according to the conformal representations discussed above. The Clifford parallels now appear as a system of mutually linked circles (and one straight line). The circles lie in a twisted fashion on a system of "nested" tori. (A torus is a surface shaped like an inner tube or the surface of a doughnut; "nested" means that these tori lie

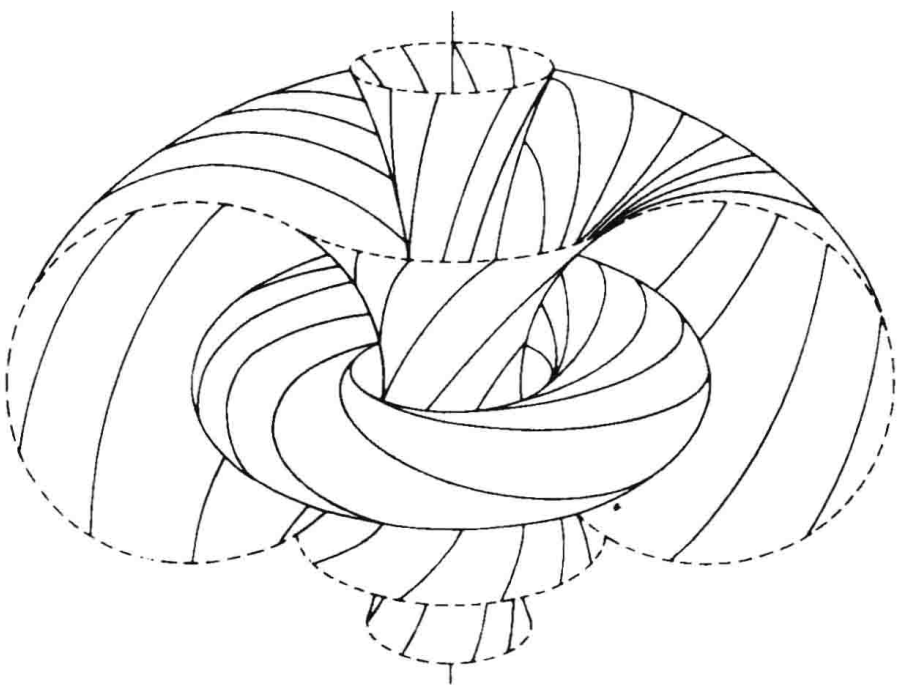


Figure 13. Clifford parallels filling the whole of S^3.

threaded within one another in continuous succession.) As can be seen, the configuration is somewhat complicated, indicating something of the intriguing possible situations that arise in higher-dimensional non-Euclidean geometry.

Irregular Geometries

So far, we have been concerned with only the simplest and most symmetrical possible non-Euclidean geometries. According to Einstein's theory, however, such models can, at best, represent only the very large scale averaged-out spatial geometry of the universe. Local irregularities in geometric structure (i.e., local "curvature") will arise on a much smaller scale, corresponding to the presence of local gravitational fields. In the immediate neighborhood of the surface of the earth this local curvature would have, as it so happens, a radius of the general order of the distance from earth to sun, or about 100,000,000 miles. (A large radius of curvature means a small curvature and *vice versa*.) As compared with the very rough value of approximately 10^{22} miles for the universe as a whole, we see that the *local* curvatures in our neighborhood dominate, locally, by a factor of something like 10^{14}. But these curvatures are still very small by any normal standards of terrestrial geometry and do not disturb the accurate employment of Euclid's geometry in the ordinary way. Even in the most extreme situation arising at the surface of a neutron star we can expect curvature radii of 100 miles or so (see Figure 14). But these small deviations from Euclidean geometry are, according to Einstein's view, *totally* responsible for the effects of gravity, so we must come to terms with them in some detail if we are to understand how gravity acts.

How, in fact, are we to give precise mathematical description to such irregular geometries, where the notions of symmetry and congruence, so basic

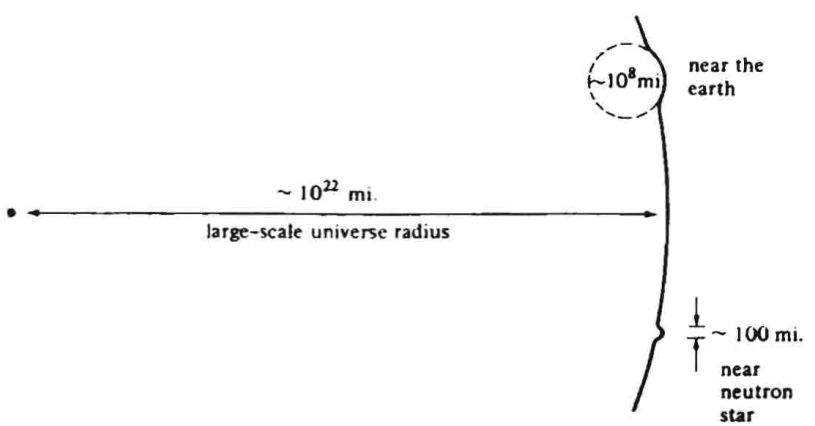


Figure 14. The rough orders of magnitude of the curvatures of our universe (not drawn to scale!).

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

to Euclidean arguments, are now totally absent? In the solution to this problem lies the essence of differential geometry.

One way that we might envisage proceeding would be to imagine the curved space that we wish to describe embedded as some kind of surface in a Euclidean space of higher dimension. This is actually a legitimate procedure and represents an older approach to differential geometry. But if, as is the case with relativity theory, we are concerned with *intrinsic* (that is, "internal" geometric) properties only, this procedure is rather indirect and is less satisfactory than certain other approaches. Any particular embedding or representation that may be employed is something extraneous to the geometry under study. We have seen this kind of thing already in our examination of L^2, S^2, and P^2, where various *different* representations in terms of E^2, or E^3, were given to describe the *same* space. It is often hard for people to grasp, at first, that such an embedding space is actually not needed. To picture a curved universe one may *feel* a need for it to be sitting inside some kind of flat background, so ingrained are the Euclidean concepts in our minds. But such an embedding space, when it is used, is for purposes of convenience only and need have no "reality" in terms of the curved geometry under consideration. (In this respect, pictures such as that given in Figure 14 can be rather misleading.)

To proceed with an intrinsic approach to differential geometry we need first the mathematical concept of a *manifold*. This generalizes the concept of a curve or surface and consists of a set of "points" together with some "local structure." The objects which are called the points of the manifold are usually undefined entities. There is no such Euclidean requirement that they be objects of infinitesimal size, devoid of physical extent. In fact we can construct examples of manifolds were the "points" themselves represent finite, extended objects. For example, each point of our manifold might represent a single location (i.e., position and orientation) of a rigid body in ordinary Euclidean 3-space, the manifold so constructed then being what is called the *configuration space* of the body (see Figure 15). Since six parameters are required to describe the location of the body (three for the position of its

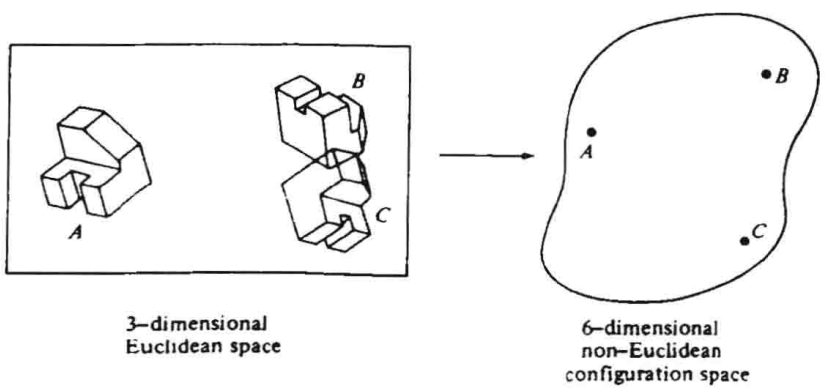


Figure 15. The configuration space of a rigid body.

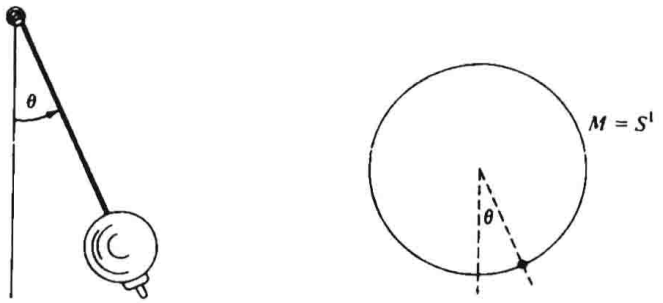


Figure 16. Configuration space of a simple pendulum.

center of mass and three for its orientation in space) the manifold we obtain will be six-dimensional (and, in fact, non-Euclidean—closely related to P^3). The word "point," when used in such a context to denote a rigid body location, for example, serves mainly to channel ones thinking in a useful direction: in the picture conjured up in ones mind's eye, such a "point" is, indeed, devoid of physical extent, unlike the original picture of the rigid body itself.

In this example it is hard to visualize the configuration space in detail, since it is six-dimensional. This does not detract from the usefulness of the configuration-space concept, however, and configuration spaces of much larger dimension are often considered. (For example, a tiny pea-sized box containing a gas with 10^{18} molecules—considered as point particles—would have a 3,000,000,000,000,000,000,000-dimensional configuration space!) But to fix ideas at this stage, let us consider a simpler situation, namely that of a rigid body (a "simple pendulum") which is constrained to move only in one plane and which is pivoted at one point, so that it has only one degree of freedom. The configuration space, in this case, is one-dimensional and is a circle S^1, the points of which may be labelled in terms of an angle θ determining the orientation of the body (Figure 16). Note that even in this simple case, the configuration space is non-Euclidean, since a one-dimensional Euclidean space is a straight line, which has a different topology from a circle.

Configuration spaces are not the only manifolds that arise naturally in mechanics. For example, velocity or momentum variables may be incorporated in addition to position variables. In the case of the constrained rigid body just considered, the angular momentum about the pivot would provide such an additional variable, and the resulting space (called a "phase space") would then be two-dimensional, having the topology of a cylinder (see Figure 17), the height along the cylinder representing the magnitude Ω of the angular momentum. Again this space is non-Euclidean (for obvious topological reasons).

Consider now the question of "local structure" for a manifold. The most primitive local structure is *topology*. Roughly speaking, the topology of a space provides a way of telling which points are very "close" to one another and which are not, or, a little more accurately, which sequences of points approach other points indefinitely closely. It does not provide any concept of an actual "distance" between points. Topology does provide us with a

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

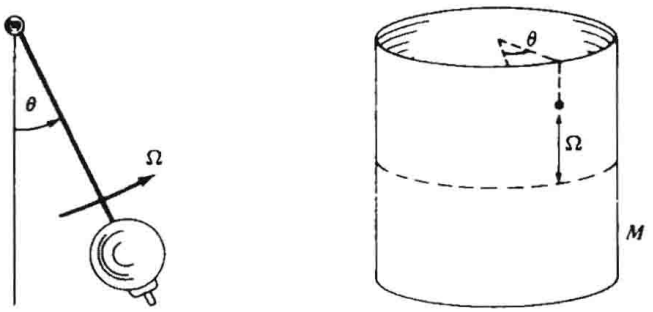


Figure 17. Phase space of a simple pendulum.

way of telling whether a curve drawn on the space is continuous (i.e., unbroken) or discontinuous, but it does not provide a distinction between a smooth curve and a kinky or angled one and it does not provide any concept of length along a curve. Topology tells us what the "continuity" structure of the space is, but not "smoothness" and certainly not "size." Thus the surface of a cube has the same topology as the surface of a sphere (S^2), though one is angled and the other smooth. But a cube or 2-sphere has a different topology from a torus (T^2) or a projective plane (P^2), or a 3-sphere (S^3), for example, or from a 2-sphere with a single point removed ($S^2 - \{\text{point}\}$), this last having the same topology as a Euclidean plane (E^2) and as a Lobachevskian plane (L^2). The cylinder of Figure 17 has the same topology as a 2-sphere with *two* points removed, i.e., as $E^2 - \{\text{point}\}$.

A topological space is called an n-dimensional (topological) *manifold* if the topology in the neighborhood of any point is the same as in an n-dimensional Euclidean space E^n. For example, L^2, S^2, P^2 are two-dimensional manifolds, while S^1 and S^3 are one- and three-dimensional manifolds, respectively. Thus, we can envisage piecing together a number of local portions of a Euclidean n-space (with regions of overlap specified) to form a manifold M; in Figure 18, this procedure is applied to produce a sphere and a torus. The rules for how this piecing together is to be achieved in any particular case

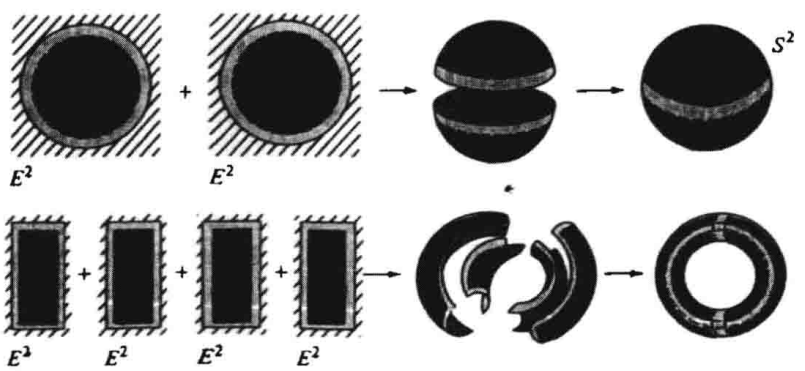


Figure 18. Piecing together a sphere and a torus out of Euclidean portions.

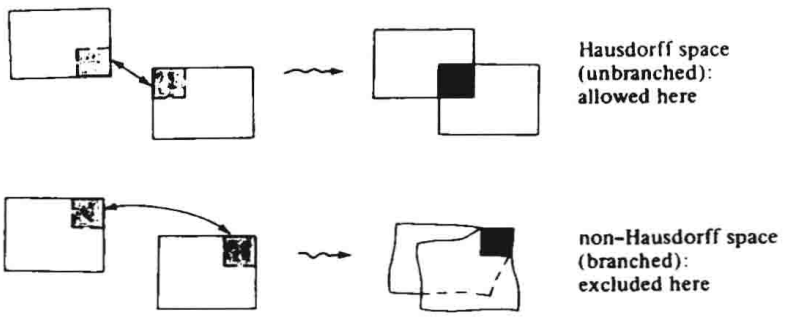


Figure 19. Hausdorff and non-Hausdorff manifolds.

can be given in an entirely intrinsic manner without reference to any embedding space: portion A of piece number 1 is to be identified with—or "glued" against—portion B of piece number 2, etc. (Hermann Weyl, 1885–1955).

(One has to be slightly careful, when specifying how these gluings are to be arranged, that the resulting space does not possess leaves or branches—that is, that the resulting space is what is called a Hausdorff manifold (see Figure 19). In fact, non-Hausdorff manifolds, in which such branching occurs, form legitimate mathematical objects worthy of study. But in normal mathematical parlance, the word "manifold," taken by itself, means a Hausdorff manifold. I shall not consider the other more general possibility in this article.)

This patching, so far, enables us to construct a manifold as a topological space only. And while we certainly do not wish to retain so rigid a structure as, say, the concept of Euclidean distance inherited from that in the individual patches, topology, by itself, is too weak a structure to support differential geometry. What we need is a *differentiable structure* for a manifold M. Such a differentiable structure may be thought of as a means of specifying whether or not a (real-valued) function on M is to be regarded as "smooth."

Let us consider an example of what "smoothness" is to mean. Take the unit sphere S^2 and choose, for example, the function h which measures the vertical height of a point on S^2 above the equator (see Figure 20a). (The value of h below the equator is then negative.) This would be an example of a smooth function of S^2. Now consider another function $|h|$, defined on S^2, which measures the *absolute* vertical distance from the equator (so the value below the equator is now positive). This function is *not* smooth along the equator, as it suddenly changes its direction of increase from that given just below to that given just above the equator (see Figure 20b). On the other hand the square $h^2(=|h|^2)$ of the vertical distance *is* smooth on S^2. This is because the change in the increase of the function below and above the equator occurs suitably gradually (see Figure 20c). To make these notions adequately precise would require entering into the intricacies of differential calculus and analysis. I do not propose to do this in any detail here, assuming that an intuitive picture of the notion of smoothness may have been supplied to the reader by the above examples. But, in fact, the whole point of assign-

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

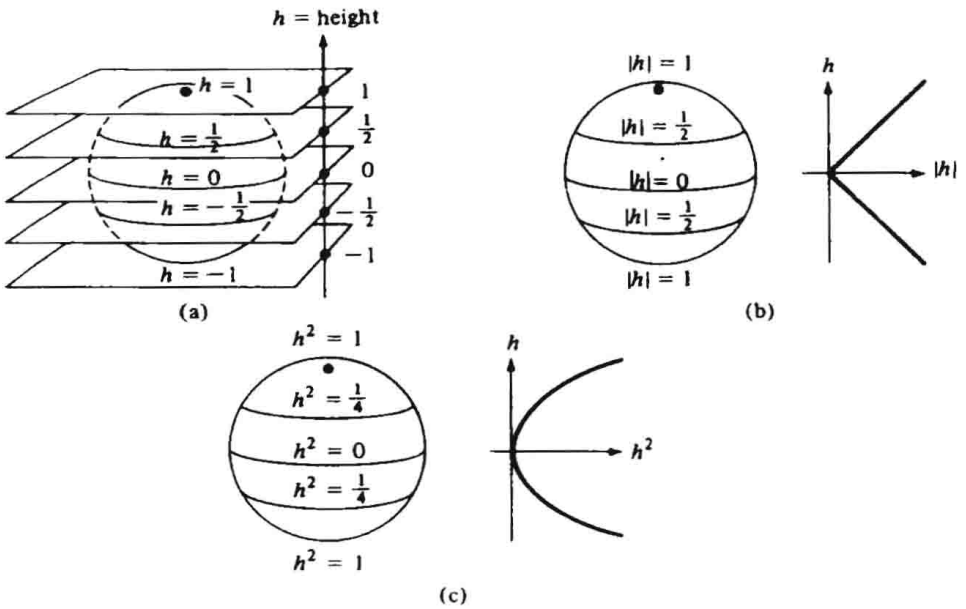


Figure 20. The smoothness of a function on S^2.

ing a differentiable structure to a manifold M is that it enables the ideas of calculus to apply on M. For this, a kind of algebra, known as tensor algebra, must be developed.

Vectors and Tensors

One way of proceeding is to try to pin things down more precisely by introducing *coordinates* into the manifold. This entails labelling, in a smooth manner, the various points of an n-dimensional manifold M with different n-tuples $(x^1, x^2, \ldots, x^n)$ of real numbers. (The indices $1,2,\ldots$ in "x^1," "x^2," ... do not here indicate that a power of x is to be taken, nor are they dimensions of spaces; they are just markers distinguishing the different coordinates.) But it may not be possible to do this all at once for the whole manifold. For example, in the cases S^2, P^2, S^3, P^3, the topology rules out such a global coordinate system. The procedure in such cases is to cover the manifold with a patchwork of overlapping coordinate systems (see Figure 21), so that each point lies in at least one patch. But some points will then lie in more than one patch and are correspondingly assigned more than one set of coordinates, so coordinate transformations must then be presented, which tell how to translate from one coordinate system to another. In this way we can provide a labelling for all points of the manifold.

This introduction of a coordinate system for M does not, however, pin things down as definitely as one might think. For the coordinate system is itself something quite arbitrary in most cases. Sometimes (as is the situation

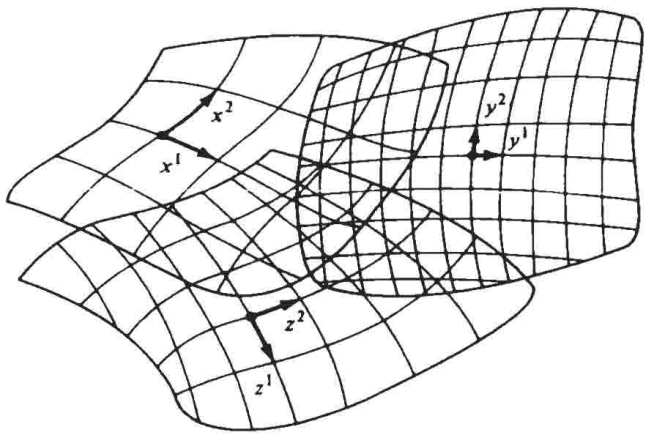


Figure 21. Overlapping local coordinate systems.

with Cartesian coordinates for the Euclidean plane) particular types of coordinates suggest themselves as more "natural" than the general curvilinear type and then it is helpful to employ such "natural" coordinates. But when M is curved in an arbitrary way, it tends to be inconvenient to try to single out special types of coordinate systems in this way. Instead, one asks for properties of the manifold M which can be ascertained equally well *whichever* system of coordinates is selected.

It may seem surprising at first that such coordinate-independent properties exist at all. But in fact they do—and the whole classical approach to differential geometry rests on this: the coordinates are introduced into the formalism only to be eliminated again, as an irrelevance, from those quantities which one is finally concerned with. There is much that is ingenious and much that is powerful in this classical approach, but it is often complicated. More recently, largely owing to the influence of the remarkable French mathematician Élie Cartan (1869–1951), a more economical coordinate-free approach has evolved, which leads directly to the required calculus of "coordinate-independent" properties. Though a detailed discussion of either the classical or coordinate-free approach is well beyond the scope of this article, it will be necessary to discuss briefly some of the key concepts that are involved.

The idea is to introduce a whole hierarchy of objects and operations, forming a kind of algebra or calculus—sometimes referred to as the *absolute differential calculus*. The most primitive objects (after the points of M themselves) are the *scalar fields*, these being simply the "smooth functions" on M that we have just been considering. A coordinate system may, if desired, be invoked to provide the initial notion of what "smoothness" means on the manifold: a function on M is then deemed to be smooth if, when expressed in terms of the coordinates, it is differentiable (in the ordinary calculus sense) as a function of these coordinates. Next, objects called *vector fields*, *covector fields*, *tensor fields*, and *connections* may be introduced in turn (and, if

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

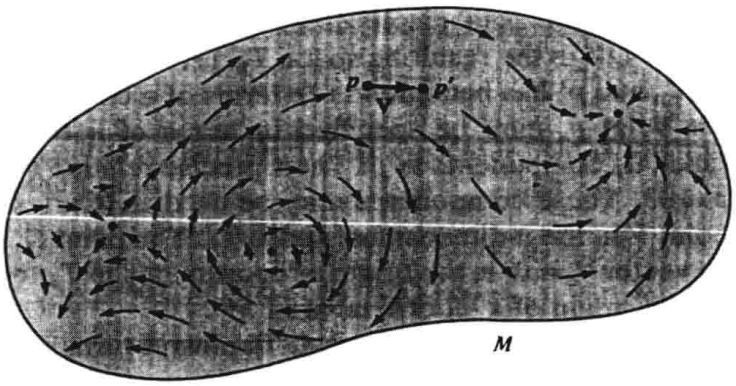


Figure 22. A vector field on M.

required, further more complicated objects too). In the classical approach, the coordinate system is called forth again and again in order to define each new class of entity in terms of the transformation properties of their explicit descriptions when the coordinate system is changed. But the economy of the coordinate-free approach is that no further mention of coordinate systems is needed once the concept of a scalar has been established.

Let us examine how this works for the next most complicated object after a scalar field, namely a *vector field*. To picture a vector field $\mathbf{V}$ (or contravariant vector field, in the older terminology) we think of a "field of arrows" on M (see Figure 22) where, crudely, each arrow is envisaged as connecting some point p of M to another one p' which lies infinitesimally away from p. In the modern viewpoint, $\mathbf{V}$ is interpreted as a *differential operator*, which is a function that acts on a scalar field ϕ to produce a new scalar field $\mathbf{V}(\phi)$, where the value of $\mathbf{V}(\phi)$ at a point p is the rate of increase of ϕ at p in the direction indicated by the arrow at p.

The set of different possible vectors (or "arrows") at a single point p of M constitutes what is called the *tangent space* T_p to M at p (see Figure 23). To picture what this tangent space means, we envisage expanding the immediate neighborhood of p in M uniformly outwards by a large amount so that

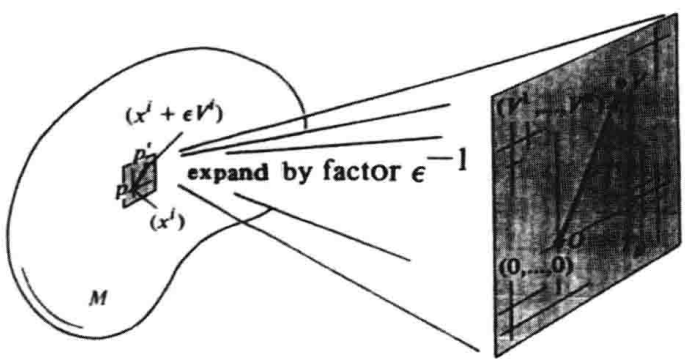


Figure 23. The tangent space T_p at a point p of M.

what used to be an infinitesimal displacement $\overrightarrow{pp'}$ on M now becomes a finite displacement $\mathbf{V} = \overrightarrow{OV}$ in the resulting space T_p. (Here p has become the origin point O of T_p and p' has become V.) In terms of a coordinate system for M, the vector $\mathbf{V}$, at p, may be specified in terms of n real numbers V^1, $V^2, \ldots, V^n$ called its components at p. Their precise relation to $\mathbf{V}$ is given in the box on p. 103. These components may be used as (Cartesian) coordinates for the point V in the n-dimensional space T_p. The point O, which represents the zero vector $\mathbf{V} = 0$, is the origin of coordinates in T_p (but, of course, p need not be the origin of x^i-coordinates in M).

Though the manifold M may be "curved," it always turns out that each T_p is, in a well-defined sense, *flat*. In fact it is a *vector space*—which means that elements of T_p can be added together (as represented geometrically by the parallelogram law) and mulitplied by numbers. However T_p should *not* be regarded (at least in the most general cases) as a Euclidean space E^n. The reason for this is that the concept of Euclidean distance between points of T_p—e.g., the distance OV which, by Pythagoras' theorem (in n-dimensions), would be given by

$$OV^2 = (V^1)^2 + (V^2)^2 + \cdots + (V^n)^2$$

—is not "invariant" under change of coordinates on M. However, it is possible to impose further structure on M which does supply such an invariant concept of Euclidean distance in each T_p. But to describe this, we must look a little further into the absolute differential calculus.

Having the concept of a vector field on M it is not hard, using a purely algebraic procedure, to pass to the next step and define a *covector field* on M. A covector field $\mathbf{A}$ (or covariant vector field, in the older terminology) may be thought of, according to the coordinate-free approach, as a certain type of function (called a *linear* function) which takes a vector field $\mathbf{V}$ to a scalar field $\mathbf{A}(\mathbf{V})$. A rough geometrical picture of a covector field $\mathbf{A}$ is a field of plane-elements (actually $(n-1)$-dimensional plane-elements) on M (see Figure 24). The plane element at p contains all the directions of all the arrows representing vectors $\mathbf{V}$ at p for which $\mathbf{A}(\mathbf{V}) = 0$. This enables us to proceed

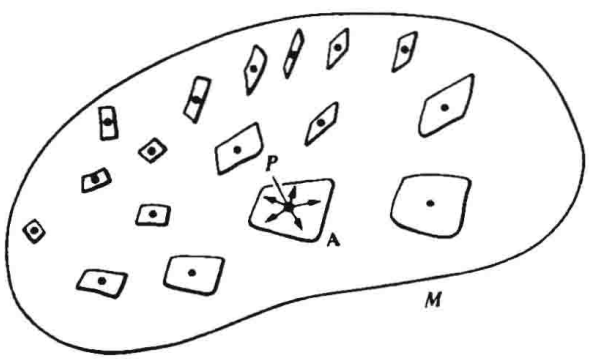


Figure 24. A covector field on M.

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

Vectors

A *vector field* **V** is a map which takes a scalar field ϕ to another scalar field $\mathbf{V}(\phi)$ such that:

$$\mathbf{V}(\phi + \psi) = \mathbf{V}(\phi) + \mathbf{V}(\psi)$$

$$\mathbf{V}(\phi\psi) = \phi\mathbf{V}(\psi) + \psi\mathbf{V}(\phi)$$

$$\mathbf{V}(\kappa) = 0 \quad \text{whenever } \kappa \text{ is constant.}$$

In terms of coordinates $x^1, x^2, \ldots, x^n$, **V** has components $V^1, V^2, \ldots, V^n$, each dependent upon $x^1, \ldots, x^n$, where

$$\mathbf{V}(\phi) = V^1\frac{\partial\phi}{\partial x^1} + V^2\frac{\partial\phi}{\partial x^2} + \cdots + V^n\frac{\partial\phi}{\partial x^n}$$

$$= V^i\frac{\partial\phi}{\partial x^i} \quad \text{using the repeated-index summation convention.}$$

to the general concept of a *tensor field* on M, which can be viewed as a similar type of map (called a *multi-linear* map) from collections of vectors and covectors to scalars (or else to vectors, covectors, or other tensors, if desired).

Perhaps the most important type of tensor field is that which defines the structure of a *Riemannian manifold*—named after the great German mathematician Bernhard Riemann (1826–1866), who, following the important work of Gauss in the two-dimensional case, virtually initiated the subject of differential geometry. This tensor field is called the *metric tensor* and is generally denoted **g**. It may be viewed as a map which takes a pair of vector fields **U**, **V** to a scalar field **g**(**U**,**V**); in terms of a coordinate system, **g** is represented as a square $n \times n$ array of components g_{ij}. (The relation between **g**(**U**,**V**) and g_{ij} is displayed in the box on p. 104, as are the two conditions of symmetry and positive-definiteness that the tensor **g** must satisfy in order to qualify as a Riemannian metric.)

The geometrical significance of **g** is that it actually does give a Euclidean metric structure (E^n) to each tangent space T_p. Recall that the Pythagorean sum-of-squares expression for OV^2 that was mentioned above will not do this because it gives a result which is dependent on the choice of coordinate system. But the expression **g**(**V**,**V**) may now be used instead:

$$OV^2 = \mathbf{g}(\mathbf{V},\mathbf{V}) = g_{ij}V^iV^j.$$

Its value is independent of coordinates simply because the quantity **g**(**V**,**W**) (in the coordinate-free approach) is something which makes no reference to a coordinate system. Furthermore, this new definition of OV can be shown

The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

Metric Tensor

$$\mathbf{g}(\mathbf{U},\mathbf{V}) = g_{11}U^1V^1 + g_{12}U^1V^2 + \cdots + g_{1n}U^1V^n$$
$$+ g_{21}U^2V^1 + g_{22}U^2V^2 + \cdots + g_{2n}U^2V^n$$
$$+ g_{n1}U^nV^1 + g_{n2}U^nV^2 + \cdots + g_{nn}U^nV^n$$
$$= g_{ij}U^iV^j \text{ using the summation convention}$$

$\mathbf{g}(\mathbf{U}, \mathbf{V}) = \mathbf{g}(\mathbf{V}, \mathbf{U})$, i.e., $g_{ij} = g_{ji}$.

$\mathbf{g}(\mathbf{V}, \mathbf{V}) > 0$ unless $\mathbf{V} = 0$, i.e., (g_{ij}) is a positive-definite matrix.

to satisfy all the properties that an ordinary Euclidean "distance from the origin" has to satisfy in a vector space. (This follows from a classical theorem which states that a coordinate system can always be chosen such that at one *particular* point p the components g_{ij} are zero for $i \neq j$ and unity for $i = j$. The expression $g_{ij}V^iV^j$ is then equal to the Pythagorean sum of squares in T_p and the required properties are satisfied. In other coordinates the expression for OV^2 would look more complicated, as is to be expected since most coordinate systems would give rise to "oblique" coordinate axes for T_p in which the Pythagorean form would not hold.)

Let me recapitulate what has been established. First there is the general concept of a manifold M, with its topology and "smoothness" structure. This is the place where the objects of absolute differential calculus can live. But at that stage no specific concept of distance is singled out, even in the tangent spaces. If we wish to introduce such a distance concept, we must do so by specifying a particular metric tensor field **g**. This enriches the structure of M and turns it into what is called a *Riemannian manifold*. The fact that the smooth manifold structure *alone* does not give us a distance concept is graphically illustrated in Escher's woodcuts (see Figure 2); the Euclidean metric on the unit disc is very different from the Riemannian ("Escher") metric of L^2, since according to the latter, the devils are all the same size!

In geometrical terms, the assignment of such a Riemannian structure means the assignment of a particular concept of *distance* between any pair of infinitesimally separated points p, p'. Hence, stringing such points together into a (smooth) curve on M we obtain a concept of distance measured along the curve, called the (Riemannian) *length* of the curve (see Figure 25). Figure 25 also illustrates another Euclidean-type concept which a Riemannian structure defines, namely the angle between two vectors, or the angle of intersection between two curves.

The "curve-length" concept on M completely defines the Riemannian

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

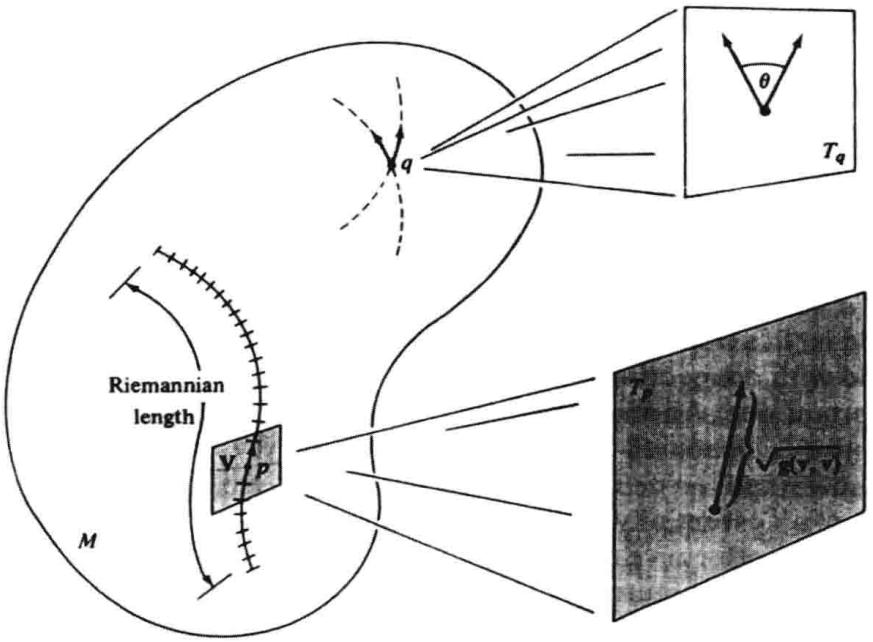


Figure 25. Lengths and angles in a Riemannian space.

structure on M: the two are equivalent. On the other hand, the "angle" concept does not quite define the Riemannian structure, but defines something weaker, known as a *conformal* structure. Recall the conformal representations of L^2 and S^2 in E^2. These are examples in which one Riemannian space (L^2 or $S^2-\{\text{point}\}$) is mapped into another (E^2) with a different Riemannian structure (so distances are not faithfully represented) but where the conformal structure is preserved in each case (so angles are faithfully represented).

Curvature

A Riemannian manifold is something that looks *locally* like a Euclidean space—in the sense that the tangent spaces are each given a Euclidean structure. Euclidean space E^n is itself a particular example of a Riemannian manifold. So also are the non-Euclidean spaces L^n, S^n, P^n, as well as the more general irregular spaces that we are concerned with now. The *deviations* from Euclidean geometry can be described, on a local scale, in terms of a certain tensor field $\mathbf{R}$ known as the (Riemann–Christoffel) *curvature* tensor. The geometrical meaning of the tensor $\mathbf{R}$ is best understood in terms of a certain operation called *covariant differentiation*, denoted ∇, a notion which was first introduced by Elwin Christoffel (1829–1900) and studied as an operation in its own right by Tullio Levi-Civita (1873–1941). The operation ∇ carries a tensor field Φ into another slightly more complicated one,

denoted $\nabla\Phi$, which measures the rate at which Φ is varying. (Scalars, vectors, and covectors may be regarded as special cases of tensors, so ∇ applies to them too.)

This operation ∇, as first introduced by Christoffel, is something which arises in a unique and rather remarkable way once a Riemannian structure $\mathbf{g}$ has been assigned to the manifold. But Levi-Civita showed how a ∇ could be defined and interpreted (though not uniquely) even when *no* specific metric $\mathbf{g}$ was given on the manifold. Such an assignment of a ∇ to M makes M a *manifold with connection*. By Christoffel's original result, any Riemannian manifold is, in a natural way, also a manifold with connection. But not every manifold with connection is Riemannian. That is only a special case, though an important one. Another important special case, which we discuss later, is the pseudo-Riemannian case needed for relativity.

To understand the geometric meaning of ∇ and of curvature we may think in terms of an essentially equivalent concept: *parallel propagation*. Suppose γ is some smooth curve on M, and let the vector $\mathbf{U}$ be tangential to γ all the way along it (see Figure 26). An operation $\nabla_{\mathbf{U}}$ is defined by ∇, called the propagation derivative, which measures the "rate of change" of any tensor field Φ along the curve γ. (In the special case when Φ is a scalar ϕ, this is our old friend the vector operator $\mathbf{U}$, i.e., $\nabla_{\mathbf{U}}\phi = \mathbf{U}(\phi)$.) We now define *parallel propagation* of Φ along γ by the condition $\nabla_{\mathbf{U}}\Phi = 0$, which means that the "rate of change" of Φ along γ is zero. Now consider the case when Φ is a vector $\mathbf{V}$. The parallel propagation of $\mathbf{V}$ along γ defines a system of "parallel directions" in M along γ. This is illustrated in Figure 26.

But there is a crucial respect in which this general notion of "parallel" differs from the familiar particular case of it that occurs in Euclidean space: for if γ is a *closed* loop, we may now find that a net *rotation* occurs when γ is traversed. Consider, for example, the case when M is a sphere S^2. Now parallel propagation of $\mathbf{V}$ along γ does *not* correspond to keeping $\mathbf{V}$ parallel to itself in an embedding in Euclidean space E^3 in which S^2 may be visualized as sitting, since $\mathbf{V}$ must always be kept pointing tangentially to S^2. Suppose that the vector $\mathbf{V}$ starts at a point p situated on the equator of S^2 and is moved by

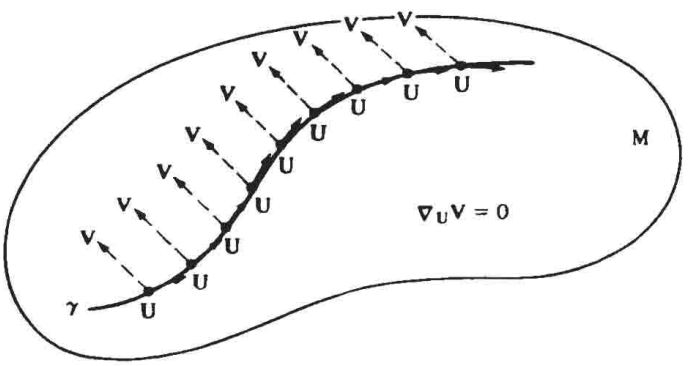


Figure 26. Parallel propagation of a vector along a curve.

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

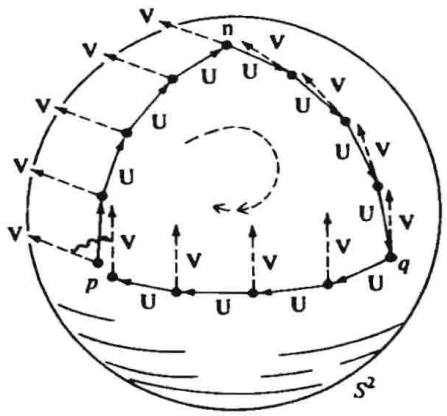


Figure 27. The effect of S^2's curvature on parallel propagation.

parallel propagation along a meridian (great circle) until it reaches the north pole (see Figure 27). If **V** begins by pointing due west, then by symmetry we can expect that it remains pointing due west until it is about to reach the north pole. Suppose it is next moved back by parallel propagation to the equator along a different meridian, in the direction opposite to the one in which **V** now points. Clearly **V** points due north all the way down until the equator is reached at q. Finally, it is returned to p by parallel propagation along the equator—pointing north all the time. But on return, we find there is a considerable discrepancy! Though the vector **V** has been moved by parallel propagation throughout, we find that it has undergone a resulting *rotation* (here through a right-angle) when it finally returns to the starting point.

This phenomenon is a direct manifestation of the presence of *curvature*, which is described quantitatively by the *curvature tensor* **R**. We may regard **R** as a map which takes three vectors **X**, **Y**, and **V** into a fourth vector $\mathbf{W} = \mathbf{R}(\mathbf{X},\mathbf{Y};\mathbf{V})$, where **W** (scaled up suitably) represents the difference between the final and initial states of a vector **V**, when **V** is carried by parallel propagation around a small parallelogram whose two sides at a point p are defined by vectors **X** and **Y** (see Figure 28). Thus, **R** measures an infinitesimal discrepancy which results from parallel propagation around an *infinites-*

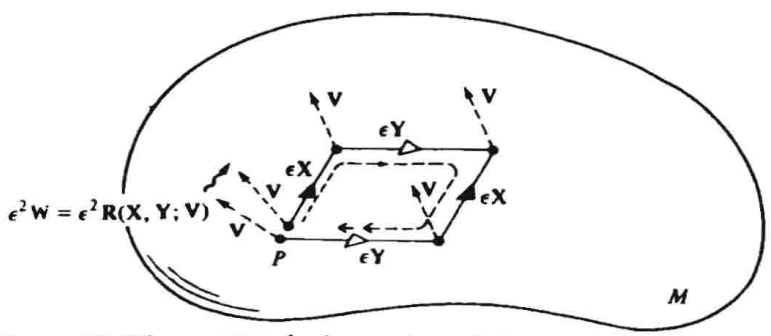


Figure 28. The geometrical meaning of the curvature tensor.

imal closed loop. The finite discrepancy for a finite loop would then be obtained by appropriately combining together these infinitesimal effects. The box on p. 109 contains details of the analytic definition of **R**.

Next, we can define *geodesics*, which are the natural analogs of straight lines for M. A geodesic is a curve with tangent vectors **U**, at different points of the curve, related to one another by parallel propagation. Hence a geodesic moves along M "parallel to itself." There is a unique (maximally extended) geodesic through a point p in each direction at p (see Figure 29).

The geodesics on S^2 are the great circles; those on each of L^2, S^2, P^2, S^3, etc. are what were referred to as the "straight lines" of those geometries. Recall also the concept of a triangle made up of such lines, and the fact that the angles summed to a quantity whose difference from π ($= 180°$) was a measure of the area of the triangle. This last fact is actually an illustration of the above-mentioned property of parallel propagation, in the special case when the curvature is constant. The integrated discrepancy, for parallel propagation of a vector around the triangle, is then just proportional to the area enclosed. Thus, as we pass around the triangle we encounter a net rotation effect which is proportional to the area of the triangle. The total angle that the sides of the triangle rotate through, as the triangle is traversed, differs from the Euclidean value by an amount proportional to this area (see Figure 29).

I have been discussing geodesics and curvature as concepts arising solely from the properties of ∇. But a concept of "geodesic" arises also directly from a metric **g**. For we could define a geodesic which connects points p and q (not too far from one another) as the curve from p to q of *minimum Riemannian length*. In fact, this definition concurs precisely with the one given in terms of ∇ provided that we choose our ∇ correctly in the first place, as the one naturally defined by **g** according to Christoffel's prescription. It is a remarkable and very basic fact of Riemannian geometry that this ∇ is in fact uniquely defined from **g** by two simple requirements. The first is that the length ($= [\mathbf{g}(\mathbf{V},\mathbf{V})]^{1/2}$) of any vector **V** remain constant under parallel propagation along any curve (or, equivalently, that $\nabla\mathbf{g} = 0$). The second is that ∇ be what is called *torsion-free* (see the box on p. 110). In geometrical terms,

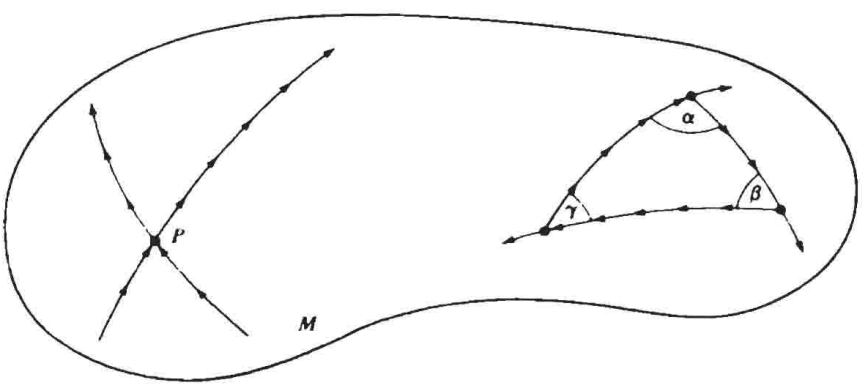


Figure 29. Geodesics and geodesic triangles.

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

Curvature Tensor

$R_{ijk}{}^{l}$ denotes the components of **r**, so $W^l = R_{ijk}{}^{l}X^iY^jV^k$ stands for $\mathbf{W} = \mathbf{R}(\mathbf{X},\mathbf{Y};\mathbf{V})$. Then

$$\mathbf{R}(\mathbf{X},\mathbf{Y};\mathbf{V}) = \{\nabla_{\mathbf{X}}\nabla_{\mathbf{Y}} - \nabla_{\mathbf{Y}}\nabla_{\mathbf{X}} - \nabla_{\mathbf{Z}}\}\mathbf{V}$$

defines **R**, where $\mathbf{Z}(\phi) = \mathbf{X}(\mathbf{Y}(\phi)) - \mathbf{Y}(\mathbf{X}(\phi))$ for scalar field ϕ.

the condition that ∇ be torsion-free is, essentially, that infinitesimal parallelograms effectively "close" (see Figure 30). This point was glossed over in the geometrical interpretation of curvature just given, but that discussion is not seriously affected by it. Though the torsion-free condition is a natural one to impose on a choice of ∇, it may be remarked that there are interesting circumstances in which it is desirable *not* to do so. For example, the concept of "parallel" that gives rise to the Clifford parallels on S^3 that were described earlier (see Figures 12 and 13) is something which arises from a ∇ *with* torsion.

Space-time

Riemannian geometry is one of the most important parts of differential geometry, but it is only a part. There are many other types of local structure that can be defined and which give rise to rich and interesting alternative geometries. One such alternative—though differing only minimally from the Riemannian case—is the *pseudo-Riemannian* geometry of Einstein's general relativity.

To appreciate the nature of this particular geometry, recall, first, some of the basic tenets of special relativity. Space and time are combined together to give a 4-dimensional picture of the world—the *space-time* of Minkowski. (Hermann Minkowski (1864–1909), a Russian-born mathematician, had lec-

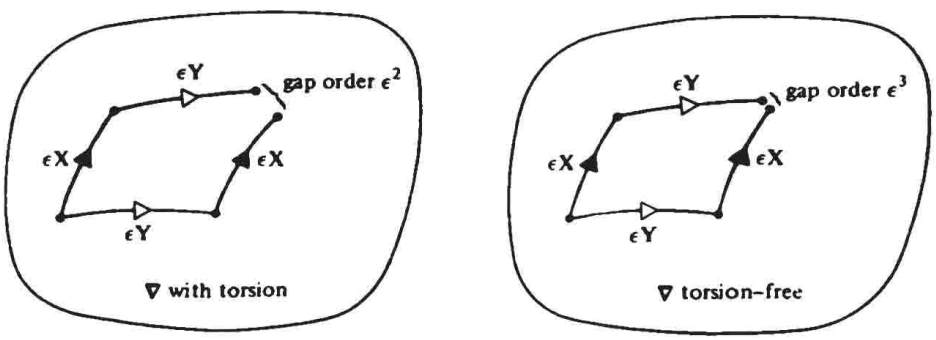


Figure 30. The torsion-free condition.

Torsion-Free Connection

∇_i denotes the components of ∇. ∇ is torsion-free if $\nabla_i\nabla_j\phi = \nabla_j\nabla_i\phi$ (ϕ a scalar field); equivalently, $\nabla_{\mathbf{U}}\mathbf{V}(\phi) - \nabla_{\mathbf{V}}\mathbf{U}(\phi) = \mathbf{U}(\mathbf{V}(\phi)) - \mathbf{V}(\mathbf{U}(\phi))$. In this case, $R_{ijk}{}^{l}V^{k} = (\nabla_i\nabla_j - \nabla_j\nabla_i)V^{l}$.

tured to the student Einstein while at the Zürich Polytechnic in Switzerland. Later, it was he who put forward the geometric space-time view of Einstein's special relativity.) This picture may be conveniently described in terms of four global coordinates t, x, y, z, where t is an ordinary time coordinate and x, y, z are ordinary Cartesian space coordinates. It simplifies matters to choose units in such a way that the speed of light (conventionally denoted c) is chosen to be equal to unity ($c = 1$). This means that units of length and time become convertible into one another according to the scheme:

$$1 \text{ second} = 299{,}792{,}458 \text{ metres} = 186{,}284 \text{ miles}$$

$$1 \text{ year} = 1 \text{ light year}.$$

Then the *light cone* of the origin is the locus

$$t^2 = x^2 + y^2 + z^2 .$$

It has two portions, namely the *future* light cone (given when $t > 0$), representing the history of a spherical light pulse travelling outwards from an initial flash at the center at $t = 0$, and the *past* light cone ($t < 0$) which is the history of a spherical light pulse converging inwards to the center at $t = 0$ (see Figure 31).

We may think of light to be propagated as particles, called photons; then the *generators* of the light cone (i.e., straight lines on the cone through the origin) represent the histories of individual photons of the light flash. The same picture would also hold for any other type of "massless" particle since all such particles travel with the speed of light. A massive particle, however, must always travel with a speed less than that of light, so the history of a free massive particle emitted in the same explosive event that produced the photons would be a straight line from the origin O lying in the interior of the future light cone (see Figure 32).

Any other particle in free (unaccelerated) motion would be likewise depicted as a straight line in Minkowski space. For a massless particle the line would be tilted at 45° to the t-axis; for a massive particle, the tilt would be less than 45°. A massive particle not in free motion would be described as a curve—the *world-line* of the particle—whose tangent vectors are every-

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

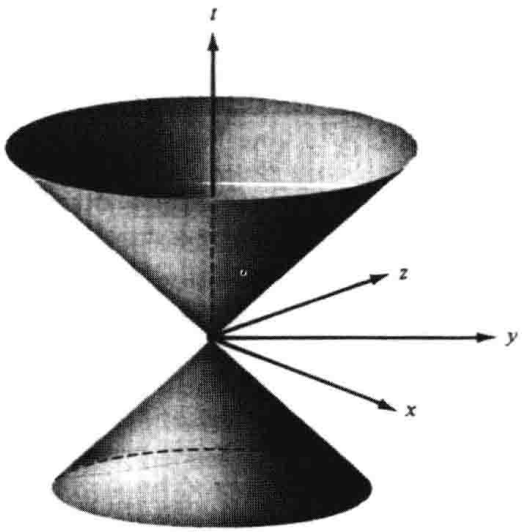


Figure 31. The light cone of the origin in Minkowski space-time.

where tilted at less than 45° to the t-axis, i.e., which everywhere lie inside the local light cone (see Figure 32). Such a world-line is called *timelike*.

Minkowski space M^4 has a natural geometry which is *pseudo-Euclidean* rather than Euclidean. The *Minkowskian distance* OQ of a point Q, with coordinates t, x, y, z, from the origin is given by

$$OQ^2 = t^2 - x^2 - y^2 - z^2 ,$$

which differs from the Pythagorean expression (for Euclidean 4-space E^4) by the presence of minus signs. (Sometimes the form $x^2 + y^2 + z^2 - t^2$ is used, but this would not conform with the particular interpretation given here.) Notice that $OQ^2 = 0$ if Q lies on the light cone of O and $OQ^2 > 0$ if Q lies inside it. Now the significance of this "distance" OQ is that (for Q within the light cone of O) it measures the time-interval experienced, between the

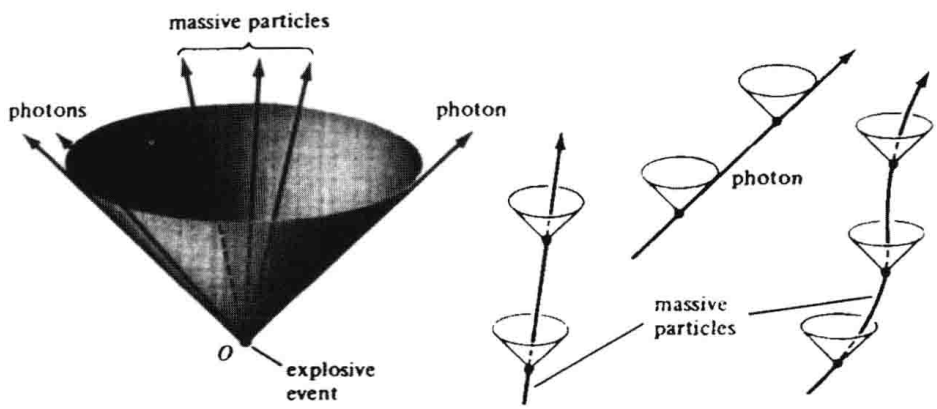


Figure 32. The world-lines of particles in Minkowski space-time.

events O and Q, by a clock whose world-line is the straight path OQ. There is a corresponding result for world-lines not passing through O. If R has coordinates t', x', y', z', then the Minkowski distance RQ is given by

$$RQ^2 = (t - t')^2 - (x - x')^2 - (y - y')^2 - (z - z')^2.$$

We have $RQ^2 > 0$ if Q lies within the light cone of R, and RQ then measures the time-interval experienced by a clock in free motion from R to Q.

The fact that measurement of time accords with such an expression, rather than with the Newtonian "absolute" time difference $t - t'$, is the key to special relativity. Time is the "distance" measure of Minkowskian geometry. And like ordinary distance in Euclidian geometry, the time measure along a timelike curve (world-line) becomes a *path-dependent* concept. If R and Q are two space-time points connected by several timelike curves (see Figure 33) then the time interval between R and Q will differ, in general, along the different curves.

Though this seems strange at first, it should be emphasized that there is overwhelming experimental support for such nonabsoluteness of time measurements (e.g., lifetimes of cosmic ray particles created in the upper atmosphere, accurate time measurements made in airplanes, the behavior of particles in high energy accelerators). We are familiar with the fact that ordinary Euclidean distance is a path-dependent concept. But our intuitions about time are built up from experiences in which the path-dependence effect is very minute—because ordinary velocities are so much less than that of light.

In Euclidean geometry, the "straight" path RQ would be the one along which the distance measure is a minimum. There is a curious reversal of this

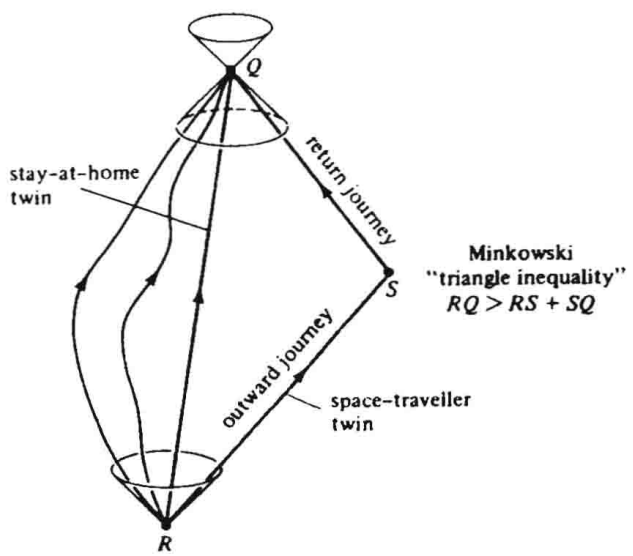


Figure 33. The "clock paradox."

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

for Minkowskian geometry: among all time-like curves from R to Q the straight (unaccelerated) world-line has *maximum* length (i.e., time interval). This relates to the famous so-called clock paradox of special relativity (see Figure 33). Two twins start from R, one of whom remains unaccelerated on his home planet, while the other travels outwards at a sizable fraction of the speed of light to a distant planet at S, whereupon he turns back for home. When they are reunited at Q, it is found that the traveller has aged less than the stay-at-home! The stay-at-home's time which elapsed during the journey of his brother is just RQ—and this must be *greater* than that measured along the traveller's nonstraight world-line.

It is worth remarking at this juncture that the Minkowskian "unit sphere," given by the set of points Q at unit Minkowskian distance from the origin O ($OQ = 1$), is actually a Lobachevskian space L^3 (or, rather two copies of such a space). The situation is depicted in Figure 34 in the case of two space-dimensions. (The conformal and projective representations of L^3 that were described earlier can be obtained by stereographic projection from $O' = (-1, 0, 0, 0)$ to the E^3 hyperplane $t = 0$, and from $O = (0, 0, 0, 0)$ to $t = 1$, respectively. This is in close analogy with the projections of S^3 and P^3 to E^3 illustrated in Figures 7 and 9.)

Velocities in special relativity are often represented as Minkowski (future-pointing) unit vectors, that is, by points on the Minkowskian unit sphere. Thus the space of velocities is really *non*-Euclidean, being a Lobachevskian space L^3. This relates to the fact that, in relativity, velocities do *not* add together in the familiar Galilean–Newtonian way.

We now pass to *general* relativity. The flat Minkowskian space-time M^4 is now to be replaced by a *curved* manifold M. In the small, the structure of M is to resemble Minkowski space, and we need a local structure which expresses this fact. So we choose a "metric" tensor $\mathbf{g}$ for M which bears the same relation to Minkowskian geometry as does a Riemannian metric tensor to Euclidean geometry. The only difference from the Riemannian case is that the positive-definiteness requirement ($\mathbf{g}(\mathbf{V}, \mathbf{V}) > 0$ unless $\mathbf{V} = 0$) is now

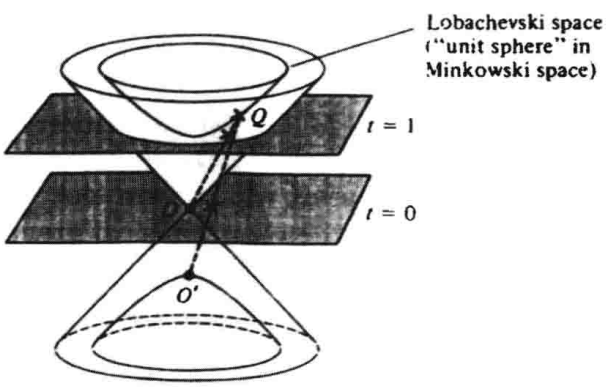


Figure 34. The Minkowsian unit sphere is a Lobachevski space.

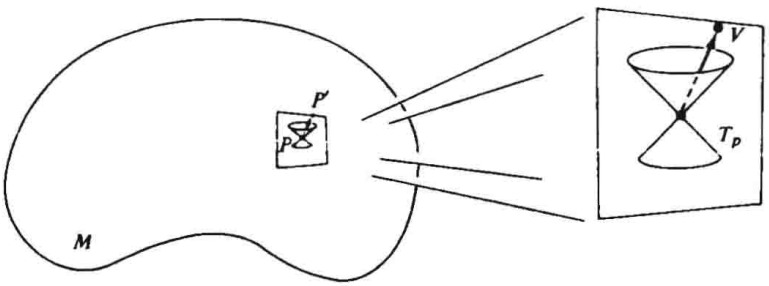


Figure 35. A curved space-time has a locally Minkowskian metric.

dropped and replaced, suitably, so that each tangent space T_p now acquires a Minkowskian rather than a Euclidean structure. This means that, at any point p, a coordinate system can be found so that, in T_p, the component form of $OV^2 = \mathbf{g}(\mathbf{V}, \mathbf{V})$ reduces to the standard Minkowskian expression $(V^0)^2 - (V^1)^2 - (V^2)^2 - (V^3)^2$. The vanishing of this expression (i.e., of the coordinate-independent expression $\mathbf{g}(\mathbf{V}, \mathbf{V})$) defines a "light cone" in each tangent space T_p—often called the *null cone* at p (see Figure 35). Such a $\mathbf{g}$ defines a pseudo-Riemannian structure for M of the kind referred to as *Lorentzian* (after the Dutch physicist Hendrik Antoon Lorentz (1853–1928)).

Vectors for which $\mathbf{g}(\mathbf{V}, \mathbf{V}) > 0$ are called *timelike* and are possible tangent vectors for world-lines of massive particles; vectors for which $\mathbf{g}(\mathbf{V}, \mathbf{V}) = 0$ are called *null* and can be tangents to world-lines of photons; those for which $\mathbf{g}(\mathbf{V}, \mathbf{V}) < 0$ are *spacelike* and are not permitted as tangents to physical particles. As in a Riemannian space a measure of length is defined, but now along timelike or null curves only. This "length" measures the time experienced by a particle having the curve as its world-line.

The null cones at the various points of the spacetime M provide the most important part of its structure. (They are defined by the metric $\mathbf{g}$ up to proportionality and constitute, in effect, the conformal structure of M.) The main physical importance of the null cones is that they determine the *causal relations* between space-time points. If a point p can be connected to a point q by a future-directed null or timelike curve then it is possible for a signal to get from p to q, but not otherwise. And it is precisely the null cone structure that determines which these curves are.

In a general Lorentzian manifold M, the causal relations may differ greatly from the situation in Minkowski space (see Figure 36). It is even possible to construct models with timelike curves which form closed loops—but such models are normally excluded on the clear physical ground that in such a universe it would in principle be possible for an astronaut to travel (or send signals) into his own past, thereby leading to the possibility of a paradoxical situation.

An example of a space-time model with unusual and interesting (though *not* paradoxical) causal properties is that representing gravitational collapse

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

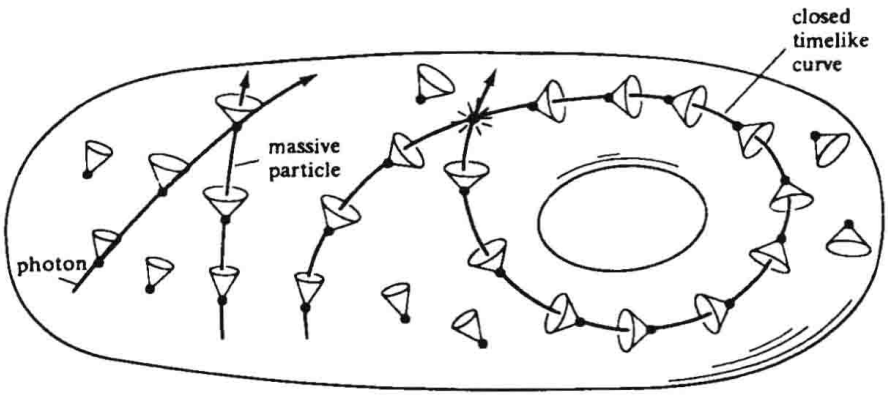


Figure 36. Null cones in a Lorentzian manifold (with a closed timelike curve).

to a *black hole* (see Figure 37). Physical theory predicts that the end-point of the evolution of a massive star—of, say, twenty times the mass of our own sun—would be such a black hole. The most noteworthy feature of this model is the presence of a space-time 3-surface, called the *absolute event horizon*, with the property that signals or particles emitted inside it cannot escape to the outside world. Signals or particles can, however, get from out-

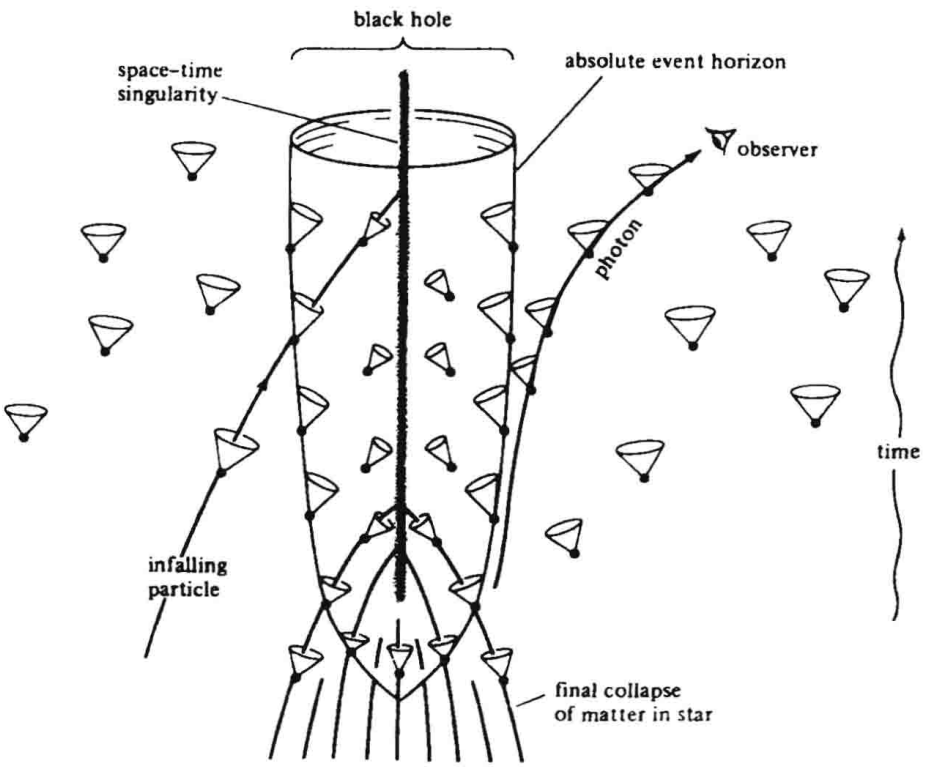


Figure 37. Collapse to a black hole.

side the horizon to inside it. The matter which collapsed to form the black hole can itself not communicate signals to the outside, once it has fallen within the horizon. Instead, this matter falls inwards to the vicinity of the center, at which it encounters a *space-time singularity* where the magnitude of the curvature of space-time approaches infinity!

Einstein's Equations

The reader may reasonably ask, however, why he should accept that space-time is curved at all, let alone why it should be curved to the extreme degree encountered in the black-hole model. In fact small deviations from the flat Minkowskian form of the metric have been experimentally detected, these deviations occurring by virtue of the presence of a gravitational field. In the 1960 experiment of Pound and Rebka, the ratio between clock rates at the top and the bottom of a 22.6 meter tower was measured and found to be the value 1.0000000000000025 that had been predicted by general relativity. Since clock rates measure space-time geometry, this may be taken as a direct measurement of the deviation of the actual geometry of space-time from the Minkowskian form. There is now a host of rather more conclusive tests of the theory available, and the theory must now be regarded as very well confirmed. For example, a fairly direct test of the non-Minkowskian nature of the null cone structure is afforded by the phenomenon of bending of light by the gravitational field of the sun—since light rays (in vacuum) follow the null cones (Figure 38). The bending effect was observed by Sir Arthur Eddington's expedition in 1919; but much more accurate measurements with radio signals have been made quite recently and conform closely to the predictions of general relativity.

These tests all measure rather minute deviations from Minkowskian space-time geometry. Large-scale deviations could occur either in a cosmological context or else with extreme collapse situations such as in the case of a black hole. It is very probable that a black hole is responsible for the X-ray source Cygnus X-1, the X-rays being emitted when matter is dragged into the hole. So it seems that such extreme situations do actually occur on a comparatively local scale. (The black-hole horizon in Cygnus X-1 would be something of the order of 40 miles across and its distance away, about 8,000 light years.) But the expected detailed properties of such a black hole depend heavily on the exact form of the equations of general relativity. These will be considered next.

So far, the only structure of the space-time that we have looked at is its metric $\mathbf{g}$. But $\mathbf{g}$ determines a unique torsion-free ∇ (with $\nabla\mathbf{g} = 0$) in the Lorentzian case just as well as in the Riemannian one. Thus we have a notion of space-time geodesic defined, and also a curvature tensor $\mathbf{R}$. Analogously to the Riemannian situation, timelike curves that are (locally) of maximal length turn out to be geodesics with respect to this ∇. And it is a

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

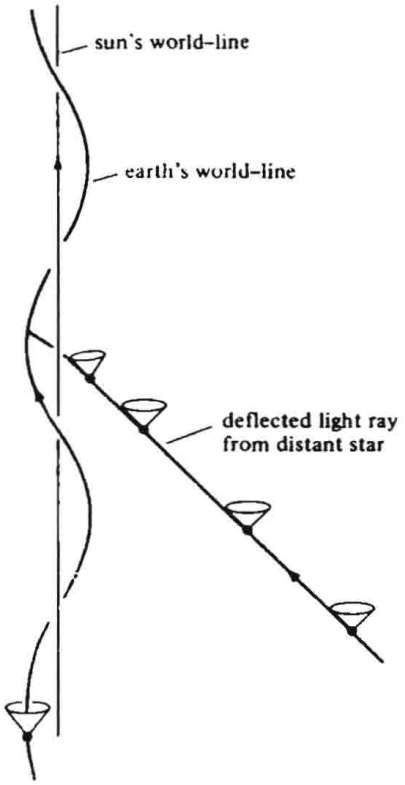


Figure 38. The earth's geodesic world-line about the sun.

hypothesis of the theory that massive test particles falling freely under gravity have world-lines which are timelike geodesics; and correspondingly, light rays are null geodesics. (Later on, one finds that it is not necessary to postulate this hypothesis; it turns out to be a consequence of the Einstein field equations.) A physical basis of this geodesic hypothesis is Einstein's *principle of equivalence*. This is essentially the observation of Galileo Galilei (1564–1642) that all bodies, whatever their constitution, fall at the same speed (neglecting air-resistance). Thus the motion of a body under free fall is determined once its initial position and velocity have been fixed, no parameters defining its mass or internal constitution being required. This is necessary for the geodesic hypothesis since a geodesic is correspondingly determined once its initial point and space-time direction at that point (i.e., spatial velocity) are known. This has the effect that freely falling objects have no local way to detect "gravitational force"—now a familiar phenomenon of the space age. Astronauts in orbit (examples of free fall under gravity) feel no gravitational force even though the intensity of the earth's gravitational field in their vicinity may differ but little from that at the earth's surface. The astronauts, together with their space-craft, are following a geodesic path in space-time. It is *we* who feel a "force," corresponding to the fact that our space-time paths are *not* geodesics. (The earth's surface is continually "pushing" us away from a geodesic, i.e., from free-fall orbit). On the

other hand, the earth itself can be regarded as a "particle" with a geodesic world-line about the sun (see Figure 38). The same applies to any other small body in orbit. None of them "feels" any gravitational force.

How is it, then, that gravitation can be detected at all, if its effects can always be locally eliminated by free fall? The answer lies in the phenomenon of *geodesic deviation*. For it is not quite true that (extended) bodies in free fall cannot locally detect the effects of gravity. There are slight differences between the motions of neighboring particles, and these relative motions *can* be detected locally. These small effects can accumulate over large volumes and become very substantial.

Imagine, first, an observer falling freely in the earth's field, where he starts off surrounded by a small sphere of particles initially at rest with respect to him. The particles nearer to the earth than himself will accelerate towards the earth slightly faster than he will; those farther from the earth, slightly more slowly; and those at the same distance, slightly inwards towards the earth's center (see Figure 39). Thus, relative to the observer, there is a slight distortion effect and the sphere accelerates unevenly about him, becoming deformed essentially into an ellipsoid. This is the *tidal effect* of gravity. (Replace the earth by the moon and our sphere by the earth, and we derive the standard explanation of the tides.) Now, if we just work within Newtonian theory, a remarkable feature of the inverse-square law shows up here: the volume of the ellipsoid is initially equal to that of the original sphere. In fact, this property characterizes the inverse-square law completely! The effect assumes that no gravitating matter is enclosed by the sphere. If matter *is* enclosed, then the volume is initially slightly reduced—by an amount proportional to the total mass of this matter. If we change our picture and imagine the "observer" at the center of the earth and the sphere of particles to be

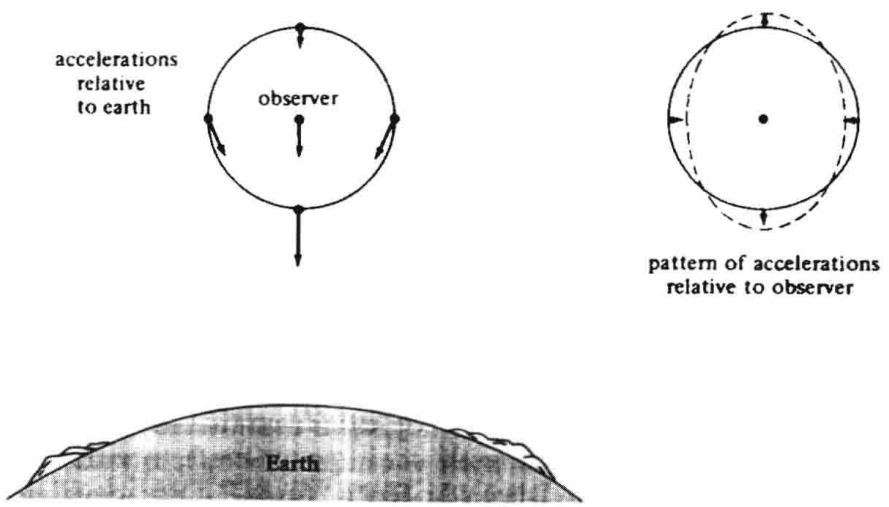


Figure 39. The distortion of a spherical volume in free fall towards earth.

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

the earth's surface, then we obtain the accumulative effect that there is acceleration inwards at every point of the sphere, leading to the familiar effect of gravity that we all feel! The volume effect is, in fact, a way of stating Newton's law of gravity.

Let us now pass to Einstein's theory. It turns out that if we state things in the above way, there is almost no change whatever from the Newtonian result. The only essentially new feature is that the volume change now is not just proportional to the mass surrounded (in the sense of mass-energy) but to the mass plus the sum of the pressures (in three perpendicular directions). But this extra effect due to the pressures is absolutely minute in all "normal" circumstances. It can only become significant when the gravitating matter becomes "relativistic," i.e., when the internal motions approach the speed of light. Such circumstances can actually begin to arise in a star which is in an extreme situation, near collapse. Then instead of tending to halt collapse, extra pressure simply serves to make matters worse owing to the increased gravitational pull. Effects such as this can contribute to the final demise of a star towards a black-hole state.

But what has this to do with space-time geometry? Let us examine our observer and his surrounding particles from the space-time point of view. We have, now, a bundle of timelike geodesics which start out parallel to one another, but where distortions are introduced as we move along the geodesics. The surrounding geodesics start by being equidistant from the central ray, but after a short while the relative accelerations between the geodesics begin to show their effects (see Figure 40). There is a formula—called the *Jacobi equation*, or *geodesic deviation equation*—which relates this relative acceleration directly to the curvature tensor. If **T** is a vector tangential to the geodesics and **V** connects a point on the central ray to a

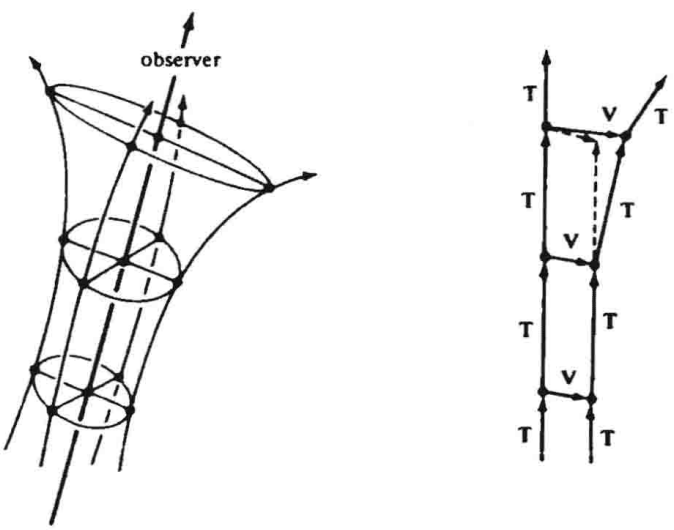


Figure 40. Space-time picture of tidal distortion.

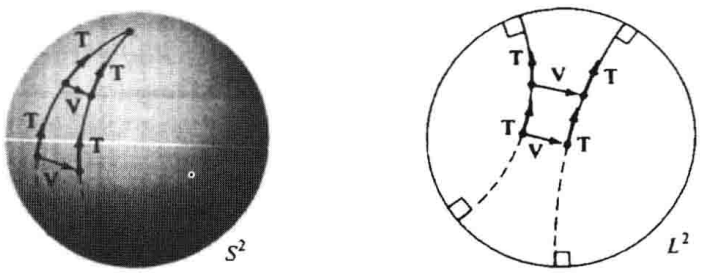


Figure 41. Geodesic deviation in S^2 and L^2.

neighboring one, then the acceleration of **V** is $(\nabla_{\mathbf{T}})^2\mathbf{V} = \mathbf{R}(\mathbf{T}, \mathbf{V}; \mathbf{T})$. In fact this result is closely related to the one depicted in Figure 28, from which it may be derived. One can also see directly this effect taking place in the geometries L^2 and S^2 that we considered at the beginning. Two geodesics which start out parallel to one another cease to be so as they are followed along for a short while (see Figure 41).

The Jacobi equation tells us that tidal distortion is a direct measurement of space-time curvature. And when we incorporate into this the statement about volumes described above, we can quite quickly obtain *Einstein's field equations*. These state the equality between two tensors, the first being a geometrical curvature quantity called the Ricci tensor, which measures the "volume change" part of the geodesic deviation, and the second describing the matter density ("mass-energy" plus "sum of pressures") being effectively the so-called *energy-momentum tensor*. (Einstein's equations are illustrated in the box on p. 121.)

The discussion of black holes is based mainly on the Einstein vacuum equations (given when the matter density is zero). When one considers cosmological models, however, it is necessary to come to terms with the structure of the space-time where matter *is* present. For this we represent a smoothed out average of the matter density as a suitable approximation to the actual contents of the universe. Local velocities are generally small (compared with that of light) and pressures are likewise small; consequently an adequate approximation is obtained by inserting the so-called energy-momentum tensor of "dust" into Einstein's equations. Furthermore, there is strong evidence that, on a large enough scale, the universe is essentially *isotropic* (that is, rotationally symmetric) about the earth's location. If, in addition, it is assumed that the earth does not occupy any special privileged position with respect to the universe as a whole, then it follows that space-time must accord with one of the famous *Friedmann models* of the universe, named after their discoverer, the Russian Alexander Alexandrovitch Friedmann (1888–1925).

There are three essentially different Friedmann models, distinguished by the value of a discrete parameter k ($= \pm 1$ or 0) which describes the spatial curvature. In the case $k = -1$ (as of now, the observationally most favored case), there are spatial sections which are all L^3 geometries; when $k = 0$ the

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

Einstein's Equations

The argument in the text gives Einstein's equations in the form:

$$R_{ij} = -8\pi G\,(T_{ij} - \frac{1}{2}T_k{}^k g_{ij}),$$

where

$R_{ij} = R_{ikj}{}^k$ = components of Ricci tensor

G = gravitational constant

T_{ij} = components of energy-momentum tensor.

An equivalent but more familiar form is:

$$R_{ij} - \frac{1}{2}R_k{}^k g_{ij} = -8\pi G T_{ij}.$$

Sometimes an extra (very small) "cosmological term" λg_{ij} is added on the left.

sections are all E^3, and when $k = 1$ they are all S^3 (or all P^3) geometries. In each case there is an initial *singularity* at which space-time curvatures approach infinity—called the *big-bang* origin of the universe. The universe expands (according to the model) very rapidly at first, but then the expansion slows down somewhat. In the case $k = 1$, the expansion actually reverses to become a collapse, and the whole model terminates in a final singularity. When $k = 0$ or -1 the expansion continues indefinitely, though rather more rapidly in the case $k = -1$ (see Figure 42).

When $k = 1$, the graph of the universe size, as plotted against time, is one branch of a cycloid (a curve traced out by a point on the circumference of a circle which here rolls on a vertical straight line). A complete cycloid is a periodic curve which repeats the same pattern again and again. This has led people to refer to the $k = 1$ model as an oscillating universe. However this is misleading because there is no way, within accepted physical laws, of getting the universe through the singularities at zero radius.

At one time it had been thought that these singularities might just be the result of the assumptions of exact spherical symmetry that have been made, since all the matter is concentrated together at one point when the singularity occurs. Perhaps if this symmetry is not required to hold exactly, people argued, the different bits of matter might "miss" one another, so that although large curvatures might be present, infinite curvatures might hopefully be avoided. This same suggestion was made as a possible means of avoiding the singularity at the center of a black hole. However, later work has shown

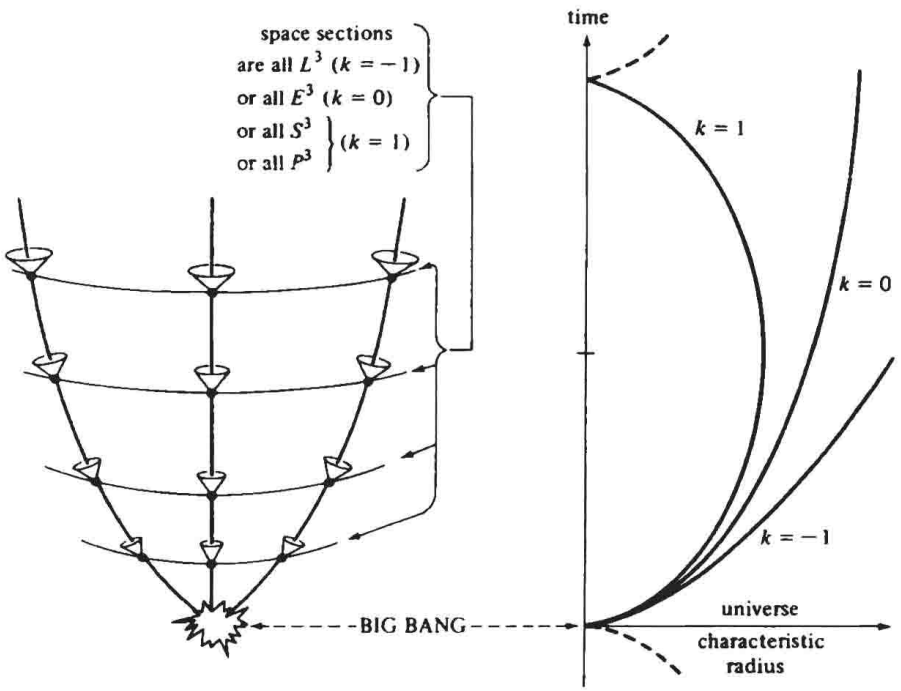


Figure 42. The Friedmann models ($\lambda = 0$ assumed).

these suggestions to be untenable. Using methods of global differential geometry it has been possible to show that the singularities are not just a feature of high symmetry but must be present in *all* general-relativistic spacetimes which qualitatively resemble the Friedmann or black-hole models. Thus, it appears that the resolution of the singularity problem is not to be found within classical general relativity. But I do not regard this as a sign of weakness in the theory since the physical conditions near singularities are not remotely like those familiar to physicists here on earth. Rather, it is the theory's strength that its essential simplicity and elegance allows arguments of such general character to be applied to it, so that rigorous deductions concerning its own limitations may be made.

New Directions

In the preceding pages, we have seen something of the power and scope of the ideas of differential geometry. But we have really witnessed only a small fraction of this power, even when restricted to applications which relate to the structure of the physical world. For example, there is another type of manifold, called a *symplectic* manifold, whose structure is somewhat like that of a (pseudo-) Riemannian manifold, but where a *skew-symmetric* tensor **s** (i.e., such that $\mathbf{s}(\mathbf{U}, \mathbf{V}) = -\mathbf{s}(\mathbf{V}, \mathbf{U})$) takes the place of the symmetric tensor **g**. The phase spaces of classical systems (like that depicted in Figure 17)

 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

are always symplectic manifolds. Yet again, there are important types of manifolds called *fibre bundles*, introduced by pure mathematicians to handle problems within differential geometry itself, but now finding significant application within theoretical physics, such as to the so-called gauge theories of elementary particle interactions.

However, to me, the most exciting new applications of geometry to the structure of the world lie within the realm of *complex* geometry. Recall the definition of a complex number: $a + ib$, where a and b are ordinary real numbers and $i = \sqrt{-1}$. Such numbers may be used in place of the real numbers of ordinary real differential geometry. In this way we may define a *complex manifold*, where complex coordinates and complex "smooth" functions (called *holomorphic* functions) take the place of the previous real ones. Recall also the representation of complex numbers as points of an ordinary plane—called the Argand (or Gauss) plane (Figure 43). This shows how, in "real" terms, every complex coordinate describes *two* dimensions. Each complex n-dimensional coordinate patch is actually $2n$-dimensional in real terms, and a complex n-dimensional manifold is $2n$-dimensional as a real manifold (with a special type of local structure).

Now, the very simplest of all closed complex manifolds—a complex analogue of a real 1-dimensional closed loop—is the (conformal) sphere S^2. This is covered by two complex 1-dimensional coordinate patches (Argand planes) which may be obtained by stereographic projection from the north and south poles of S^2, respectively (recall Figure 7). Thus S^2 is a closed complex manifold of *one* dimension or a *complex closed* curve!

Now think of S^2 in a different way—as the *celestial sphere*, or *sky* of an observer in space. Suppose a second observer, momentarily coincident with the first, moves relative to the first and is rotated with respect to him. Each

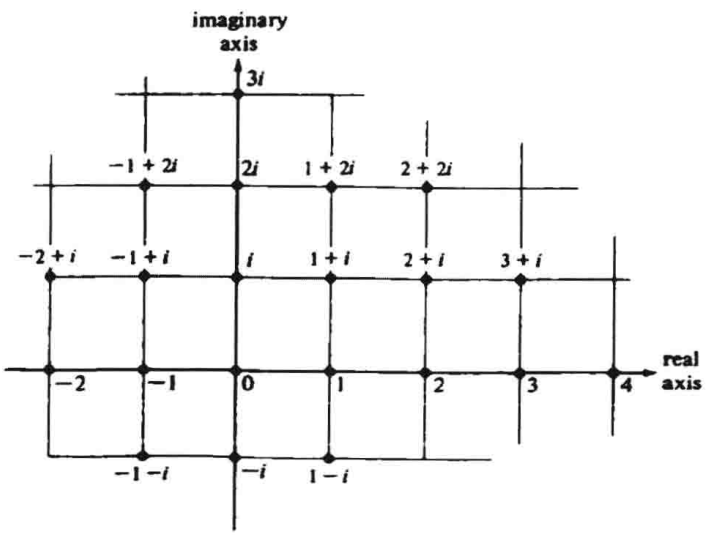


Figure 43. The Argand plane.

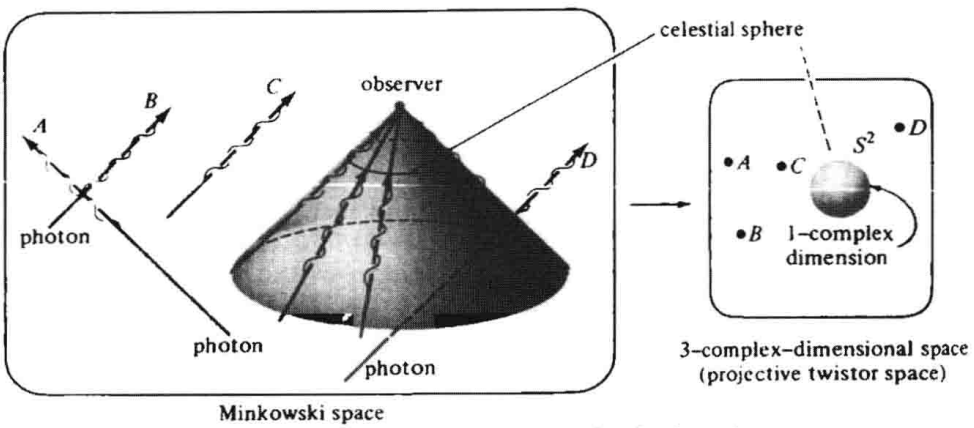


Figure 44. The complex geometry of spinning photons.

observer may use a quite different description of his "sky" in terms of complex coordinates but, remarkably, the two descriptions will give S^2 a completely equivalent structure as a "complex curve." Even more remarkably, the *Lorentz group*, whose elements represent the freedom of passing from one observer to another, corresponds precisely to the group of complex self-transformations of this complex curve. The most fundamental of symmetry groups in physics is, therefore, just the symmetry of a simple complex closed curve!

Many properties of the Lorentz group and Lorentz transformations may be very rapidly obtained using this viewpoint. But this is just the beginning. The celestial sphere may be thought of as a section through the past light cone of the observer—since the light rays entering his eye generate this cone. But the system of *all* light rays, not just those that enter his eye, can also be naturally represented as points of a complex manifold—this time 3-complex-dimensional—provided that we restrict attention to Minkowski space and that we correctly treat the energy and spin of the photons (see Figure 44). This leads us into *twistor* theory, in which complex manifolds play a key role in the description of physics.

Complex geometry is a remarkably rich, elegant and powerful area of mathematics. A great deal of intricate structure can be subsumed into logically simple concepts. For example, even the flat tangent space of a 2-complex-dimensional manifold reveals, upon examination, our now familiar friend, the system of Clifford parallels on S^3. And this, in its turn, has direct relevance to twistor theory, to the structure of spinning massless particles and to the architecture of quantum mechanics. Clearly there is much yet to be revealed in such interplay as this between mathematics and physics, or between the constructions of the mind and the works of nature.

340 The geometry of the universe. *Mathematics today, twelve informal essays.* Ed. L. A. Steen, Springer, New York, 1978

Suggestions for Further Reading

General

Kaufmann, William J. *Cosmic Frontiers of General Relativity*. Little, Brown, Boston, 1977.

Rindler, Wolfgang. *Essential Relativity*. D. van Nostrand, New York, 1969; Springer-Verlag, New York, 1977.

Taylor, Edwin F. and Wheeler, John A. *Spacetime Physics*. Freeman, San Francisco, 1963, 1966.

Technical

Guillemin, Victor and Pollack, Alan. *Differential Topology*. Prentice-Hall, Englewood Cliffs, 1974.

Hawking, S.W. and Ellis, G.F.R. *The Large Scale Structure of Space-Time*. Cambridge University Pr, New York, 1973.

Misner, Charles W., Thorne, Kip S., and Wheeler, John A. *Gravitation*. Freeman, San Francisco, 1973.

O'Neill, Barrett. *Elementary Differential Geometry*. Academic Pr, New York, 1966.

99
EINSTEIN'S VISION AND THE MATHEMATICS OF THE NATURAL WORLD

The Sciences (1979), 6–9

Einstein's Vision and the Mathematics of the Natural World

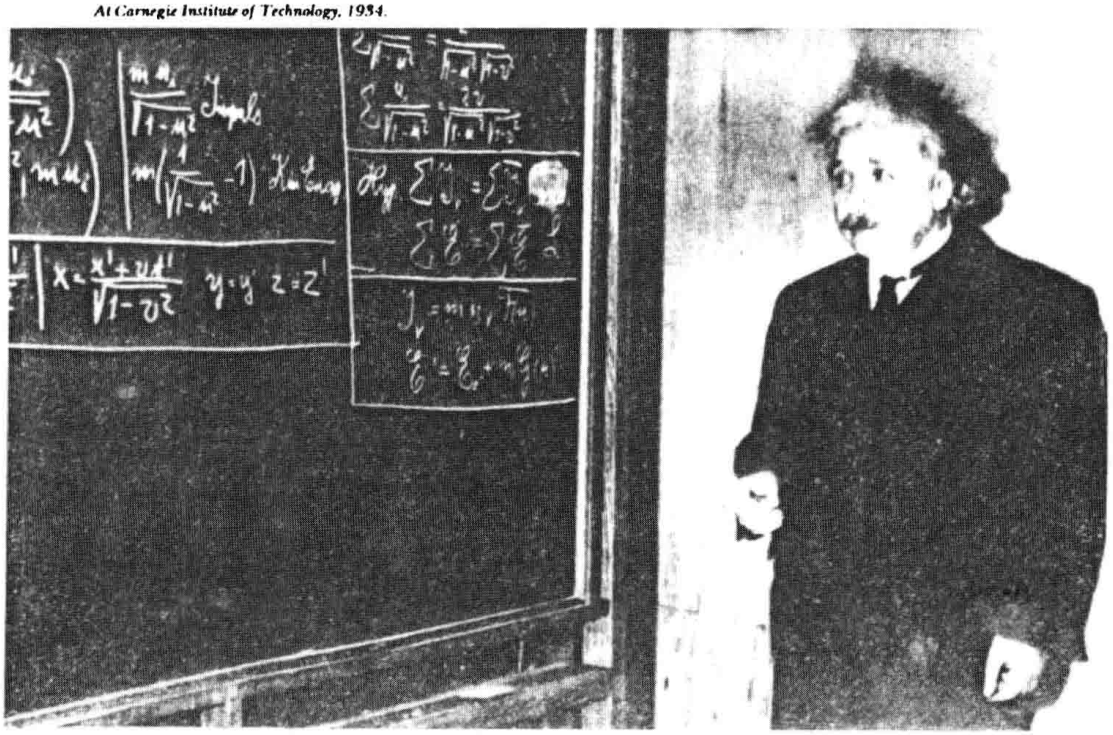
by Roger Penrose

In Einstein's young and fertile years he had apparently been somewhat skeptical about the value of sophisticated mathematics as a guide for understanding the workings of the natural world. That appreciation came to him only later; and he was moved towards it only after his own profound contributions to physics were seen to have an intimate relation to some significant developments in mathematical thought.

In Einstein's own words: ". . . it was not clear to me as a student that the approach to a more profound knowledge of the basic principles of physics is tied up with the most intricate mathematical methods. This dawned on me only gradually, after years of independent scientific work."

From an early age Einstein had been attracted to mathematics and had been impressed, particularly, by the certainty of truth that was to be found in it.

At Carnegie Institute of Technology, 1934.



He was an able enough mathematician, in fact, although he was not an outstanding one of comparable caliber to the leading mathematicians of his day, such as the highly original and broad-ranging French mathematician Henri Poincaré (who, even before Einstein, had worked out the basic mathematics of special relativity) or the brilliant but shortlived Russian-born Hermann Minkowski (who provided the space-time view of Einstein's special relativity which was invaluable for its later development into general relativity) or the masterful German mathematician David Hilbert (who had derived the field equations of general relativity independently of Einstein by means of a powerful mathematical device known as a variational principle).

Einstein's supreme abilities lay elsewhere, in being able to see right through to the heart of a physical problem, or being able to recognize in a simple and supposedly "obvious" physical fact, that a deep and unexpected truth lay hidden beneath it. Nor did he need to be driven towards his understandings by an ever-more-compelling mass of accumulated physical data. He seemed somehow to be able to sense the truth by some miraculous combination of penetrating clear thinking and an extraordinary instinct for the workings of the natural world.

But, by his own admission, these instincts did not extend, in so definitive a way, to the world of mathematics: ". . . my intuition was not strong enough in the field of mathematics in order to differentiate clearly the fundamentally important, that which is really basic, from the rest of the more or less dispensable erudition."

Yet, his physical instincts were so remarkable that he could perceive the type of theory of gravitation that he needed long before he had mastered the intricate formalism that was required for its precise mathematical description. This had been necessary in order that he should know beforehand what field of mathematics he needed to take on. It was not until 1915 that the final form of this theory, the General Theory of Relativity, appeared, but the deep physical insights that led to this theory—constituting what is perhaps the single most profoundly original contribution to scientific thought—had been already evident to Einstein seven years earlier.

Recall that the special theory was published in 1905. This special theory contained what were, at that time, many novel and counter-intuitive physical ideas: the equivalence of mass with energy, the dependence upon velocity of the very concepts of time, space and mass, the fact that the order of occurrence of two distant events may have no objective meaning. Yet the *mathematics* of special relativity was not, as such things go, excessively complicated or hard to master. It was the *physical* ideas that were so difficult to come to terms with. And when Minkowski produced his space-time interpretation—a generalization of Euclidean geometry in which time played the role of a fourth dimension—the mathematical formalism became especially transparent. (Yet, physically, many aspects of the theory remained hard to accept: such as the fact that, now, we seemed to be provided with a totally determinate and preassigned picture of the world in which past, present and future are all given as one indivisible unity!)

Symmetry Out of Chaos

The mathematics of general relativity, on the other hand, is not nearly so readily comprehended, and detailed calculations within the theory can be exceedingly complicated. Unlike the comparatively simple geometry of Minkowski, the geometry of general relativity requires the elaborate mathematical machinery of differential geometry and tensor calculus. The reason for this increase in mathematical complexity and profundity is not in itself hard to grasp. The geometry of Minkowski is completely uniform, flat, symmetrical and, essentially, featureless. In contrast, the geometry of general relativity, called *pseudo-Riemannian geometry* (after the great nineteenth century German mathematician Bernhard Riemann) is nonuniform and curved, the detailed features of the geometry ever-changing from point to point. The geometry itself is generally quite without any kind of symmetry.

Yet symmetry of a different kind now emerges out of this apparent chaos. It is a more subtle type of symmetry, residing not in the geometry but in the mathematics itself. Thus, the very object that describes the irregularity of the geometry, namely the Riemann curvature tensor, is itself an object with a symmetry and beauty of its own. The geometry of space-time is already a considerable abstraction from the physical world of our experiences. Here we must proceed one step further in abstraction if we are to perceive where the beauty and simplicity of form resides, namely within the mathematics itself. And it has taken many years of further development by numerous mathematicians, since Einstein's general theory first appeared, to improve the mathematical formalism in order to bring out this beauty and the simplicity to the full (and, indeed, an important formal step in this direction—the "summation convention"—was made by Einstein himself). The mathematical subject of differential geometry, to which all these developments pertain, has, in fact, received a very considerable impetus from Einstein's theory. And it is a subject with many applications in other areas of physics and mathemat-

Roger Penrose, who is in the Mathematical Institute at Oxford University, is also a visiting professor at the University of California at Berkeley. His field of interest is General Relativity.

ics quite aside from General Relativity.

Perhaps we are here beginning to perceive a blurring of the borderline between mathematics and physics. It is a blurring that is present even more strikingly in quantum mechanics than in relativity. The precise point at which "physical reality" leaves off and pure "mathematical formalism" takes over becomes hard to discern. An object such as a table, for example, seems as "physically real" and as tangible as an object can be. Yet, when we ask what that table is composed of, we find, increasingly as the units considered become more basic and fundamental that, at the same time, they become more abstract and mathematical. Thus, while cells or even molecules seem to be objects of tangible reality, this is progressively less clear for atoms, nuclei and for elementary particles, and it is less clear still for the *empty space* that we learn that the table is largely composed of!

What is an electron? Is it a physical object or an element of some mathematical system? Indeed, we learn from the exclusion principle of quantum mechanics that individual electrons cannot be meaningfully distinguished from one another, but only the conglomerate of all electrons in the universe can be considered as a discernible entity. Thus, while it may be the case that in appropriate circumstances a single electron may be far enough from its neighbors to be considered, in practice, as a separate entity by itself, this is only an approximation. For there are other circumstances, not differing from these in any clearly defined absolute way, in which the electrons do *not* at all behave as separate distinguishable objects and cannot correctly be treated in this way. The same is also true, in a slightly different way, of photons, the apparently individual quanta that go to make up light. While they remain separated from their fellows they behave as individuals, yet they cannot strictly accurately be treated as such. Photons are essentially indistinguishable from one another and it is not absolutely meaningful to say where one photon leaves off and the next one begins. It is a point of some historical interest, moreover, that not only was Einstein the first to ascertain the physical existence of photons but it was he also (together with the Indian physicist S. N. Bose) who first explored the statistical consequences of this physical indistinguishability.

But even this picture ignores the interactions between particles. Even the conglomerate of all electrons or of all photons cannot be considered on its own, except as an approximation. We are led, instead, to the mathematical picture presented by a fully interacting quantum field theory. But here the mathematics begins to let us down. There is, as yet, no complete, contradiction-free mathematical formalism available which can give an accurate description of all known particles with their interactions. But in particular cases, such as with the interactions just between electrons and photons, a good theory exists (called quantum electrodynamics) and some remarkable agreements between the theory and observation have been achieved.

What about space—or, more correctly, space-time? We may ask what is that? Einstein has led us to the mathematically elegant and physically extremely accurate picture as a pseudo-Riemannian geometry. Yet we know that even this picture cannot ultimately hold true. For the phenomena of quantum mechanics must also play a role in the structure of space-time itself and serve to modify in some fundamental, though so far quite unobserved way, our now familiar pseudo-Riemannian picture. This is the problem of *quantum gravity*, on which a number of leading theoretical physicists are now actively engaged. It is hoped that by achieving a deeper understanding of the structure of space-time itself, a more coherent picture of physics as a whole will emerge. And though some modification of Einstein's geometrical view is sought, such modification, if it is to have any hope at all of success, must be built upon the very one that we were given in 1915 by Einstein.

The Mathematical Physics of the Future?

But what kind of answer to all these various questions can we expect? It is hard to conceive of any answer that is not essentially a mathematical one. And mathematics itself has grown enormously since the days of Einstein's youth. It has become ever-more diversified, although some very powerful unifying concepts have been introduced. It has become more abstract while at the same time, in many areas, more complicated. Even the most erudite mathematicians of our time can now hope to master only a small fraction of the body of existing mathematical knowledge and understanding. The problem that confronted the student Einstein, of selecting what is physically important in mathematics from the rest, is now present to a greater degree than ever it was.

Yet one thing, to my mind, seems clear. The relevant mathematics will become more and more abstract and further and further removed from our immediate experiences. It may well be (as is certainly to be hoped) that there will be a greater logical simplicity in the abstract mathematics, the closer that our understanding gets to the basis of physical truth. Yet we must expect that, as the length of the chain which connects our immediate experiences to our mathematical understanding of that physical truth increases, so also will the mathematical complication which is entailed in actually deriving the consequences of such a theory increase.

But does this not seem a program of despair? If a proper comprehension of the meaning and har-

mony of this world, in which we find ourselves, is ultimately to depend upon an appreciation of a mathematics that is at the one end so abstract and at the other so complicated that even a Poincaré or a Hilbert would have difficulty coming to terms with it, then what hope is there for the ordinary man or woman who seeks solace in such comprehension? And even this presumes that understanding will eventually be forthcoming to at least some of our number. Yet I prefer to view this situation in a more optimistic way. It is often not important in mathematics (despite what we are frequently told!) that the details of an argument be followed through. It is sufficient, in many cases, just to be "told the answer." So long as *someone* has actually carried out the details of a complicated computation (and for parts of an argument, even a computer may well suffice!), this should be sufficient for the rest of us.

So the complication may not be so serious a problem as it seems. The comprehension of abstraction is, perhaps, the more difficult matter. But so long as there are able expositors about who also have the ability to comprehend this abstraction (and it is clear-headedness that is the prime prerequisite for both these abilities) then the eventual knowledge that such an abstract idea "works" should provide the additional impetus for such an understanding among a wider audience. There was an appreciable time, after all, when Einstein's Theory of General Relativity was supposed to have been understood by only a handful of people. Now it is considered to be not a particularly difficult option in many mathematics or physics degree courses.

So I feel optimistic that the quest is a far from hopeless one. The progress that has already been made over the centuries in our understanding of nature has been very remarkable indeed. And those astonishing developments that have already taken place in our present century owe more to one man than to any other: Albert Einstein. □

How to Fight McCarthyism

Einstein sent this letter, dated May 16, 1953, to William Frauenglass, a Brooklyn, N.Y. teacher who had refused to testify before a congressional committee. It is excerpted from the June 12, 1953 NEW YORK TIMES.

The problem with which the intellectuals of this country are confronted is very serious. The reactionary politicians have managed to instill suspicion of all intellectual efforts into the public by dangling before their eyes a danger from without. Having succeeded so far, they are now proceeding to suppress the freedom of teaching and to deprive of their positions all those who do not prove submissive, i.e., to starve them.

What ought the minority of intellectuals to do against this evil? Frankly, I can only see the revolutionary way of non-cooperation in the sense of Gandhi's. Every intellectual who is called before one of the committees ought to refuse to testify, i.e., he must be prepared for jail and economic ruin, in short, for the sacrifice of his personal welfare in the interest of the cultural welfare of his country.

However, this refusal to testify must not be based on the well-known subterfuge of invoking the Fifth Amendment against possible self-incrimination, but on the assertion that it is shameful for a blameless citizen to submit to such an inquisition and that this kind of inquisition violates the spirit of the Constitution.

If enough people are ready to take this grave step they will be successful. If not, then the intellectuals of this country deserve nothing better than the slavery which is intended for them.

At Princeton 1952 (Bettmann)

CHAPTER 99

100
THE TOPOLOGY OF RIDGE SYSTEMS

Ann. Hum. Genet. London **42** (1979), 435–44

In his later life my father took a professional interest in fingerprints and palm prints, and the topological issues that they raised. In this paper I paid my respects to him at a Memorial Conference held in Berlin in 1973, having been struck by the analogy between such things and some topological issues raised by the study of Weyl-curvature distortions of space-time (see Figures 8.1, 8.2, and 8.3 on pp. 231 and 232 of *Spinors and Space-Time*, Volume 2, by R. Penrose and W. Rindler, Cambridge University Press, 1986).

Ann. Hum. Genet., Lond. (1979), **42**, 435
Printed in Great Britain

The topology of ridge systems

By R. PENROSE
Mathematical Institute, 24 St Giles Oxford, OX1 3LB

Topology is a branch of mathematics, concerned with the study of properties which are unchanged by continuous deformations. The geometrical concepts of 'size' or of precise 'shape' are not relevant to topology. If an object is distorted continuously without being torn apart or glued together, then its topology does not change, even though its precise geometrical configuration may become drastically altered. Thus, a solid sphere and a solid cube are topologically indistinguishable from one another whereas a solid sphere, a ring, and a hollow sphere are all topologically distinct.

A topological classification of ridge patterns, therefore, will take no account of the normal geometrical relations such as 'distance' or 'angle'. We may imagine the ridge patterns to be drawn on a rubber glove. The glove may be stretched or distorted in any way whatsoever (provided it is not torn or glued) and the topology of the patterns will not change. But it is clear that distance and angle measurements may be drastically altered by such deformations. In Fig. 1, the circle (*a*) and the winding line (*b*) are topologically identical, whereas the open curve (*c*), although geometrically much more like a circle than (*b*), is topologically distinct from (*a*).

In many ways, a topological classification must be regarded as a rather crude one, since it ignores so much of the geometry. However, there are certain contexts in which topological ideas can prove very useful indeed. The subject has, in fact, been a very fruitful and actively studied one in modern mathematics. It also has applications in the physical sciences – sometimes of a rather unexpected character. Perhaps it is less surprising that applications in the biological sciences are to be found in some areas. When a living thing grows, it is to be expected that considerable changes in size and shape may take place continuously, whereas changes in the topological structure might be imagined to occur only at very special times in its development.

In addition to purely topological matters, there is another consideration that I feel may be important for classification schemes in biological contexts. This is the question of the *stability* of a classification. In Fig. 2 is an illustration of what is an unstable situation, if we are concerned with the entire topological pattern of lines. It will be noticed that the two patterns are topologically quite distinct from one another, since the left-hand pattern contains several closed loops, whereas that on the right consists entirely of two open strands. However, only a very small local distortion is needed in order to pass from the left-hand pattern to the right-hand one, namely a shunt upwards by one ridge of the portion to the right of the broken line. Any classification scheme which depends on tracing the connectivity of the entire pattern of lines is likely to be unstable in this sense. If the laying down of ridge patterns in a developing embryo is governed by purely *local* criteria, then the two patterns illustrated in Fig. 2 would have to be judged as being basically similar, even though they are topologically very different from one another. (If such differences in global pattern are to be found, for example, between identical twins even when the local pattern configurations are similar, then this could be viewed as lending support to a hypothesis that the laying down of the ridges is indeed a local process.)

0003-4800/79/0000-4216 $02.00 © 1979 UCL

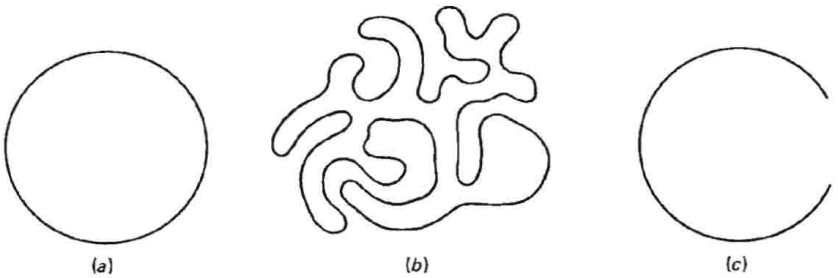


Fig. 1. The circle (*a*) and the closed curve (*b*) are topologically identical, whereas the curve (*c*) is topologically distinct from the other two.

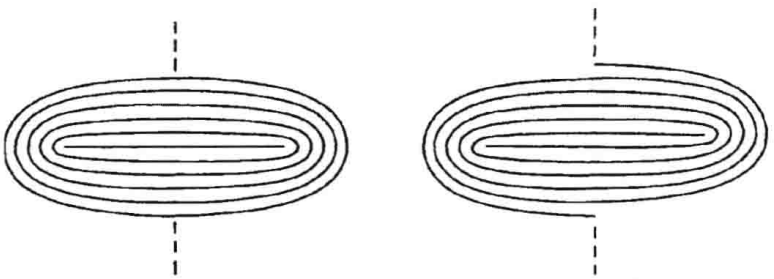


Fig. 2. The ridge pattern on the left differs very little from that on the right even though the two patterns are topologically quite different from each other.

The topological properties with which I shall be concerned here will not possess this type of instability. The ideas will also apply in other situations and it will be useful to consider ridge systems in a somewhat broader context than that which is strictly relevant to dermatoglyphics. In Fig. 3 is illustrated what is known as a *vector field*. Examples are very familiar in physics, such as a magnetic field or the velocity field of a fluid. There may be certain places where the directions of the field vectors are not defined – such as where the magnetic or velocity vector becomes zero. Such a point will be referred to as a *singularity in the direction field*. Let us assume that all the vectors are constrained to one plane (rather than three-dimensional space). Then there are two basic types of singularity which can occur, these being illustrated in Figs. 4 and 5 respectively. All other singularities may be regarded as, in a certain sense, built up from these. The distinction between the singularities of Fig. 4 and that of Fig. 5 may be seen as follows. We imagine a circle to be drawn which surrounds the singularity. As we traverse this circle in the positive sense (i.e. anticlockwise) we observe that the vector directions at the points of the circle also rotate. In the case of Fig. 4 they rotate once (i.e. through 360°) in an absolute sense positively (i.e. anticlockwise) as we return to the starting position, whereas in Fig. 5 they rotate once negatively (clockwise). Thus the singularities depicted in Fig. 4 count as positive and may be assigned a count of $+1$. Correspondingly, the singularity of Fig. 5 counts as a negative singularity and is assigned the count of -1. The same idea may be applied to more complicated singularities such as the 'dipole' field depicted in Fig. 6. Here the count is seen to be $+2$ (i.e. the vectors rotate through 720° as the singularity is circumnavigated) so we may consider it as a juxtaposition of two singularities of the type illustrated in Fig. 4. Indeed, by a slight local distortion of the vector field we may obtain two separated singularities each of which is as in Fig. 4. In the same way a count (which is some integer) may be assigned to any localized

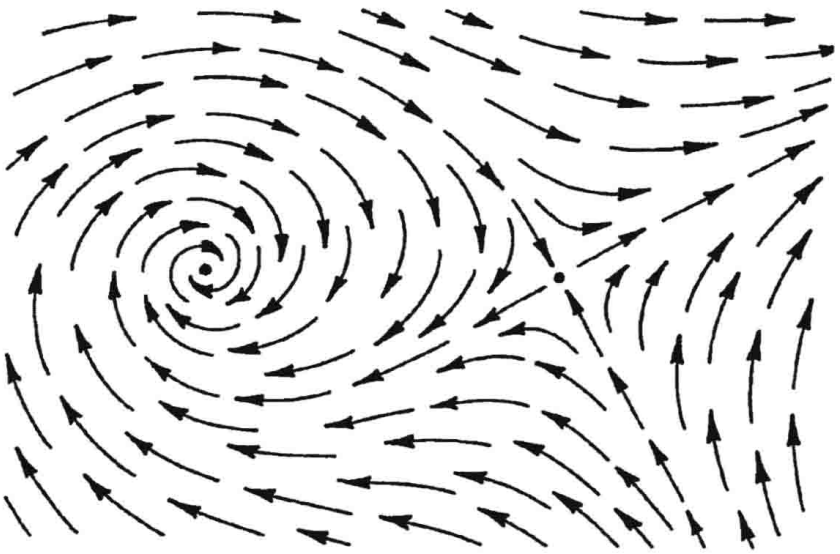


Fig. 3. A vector field.

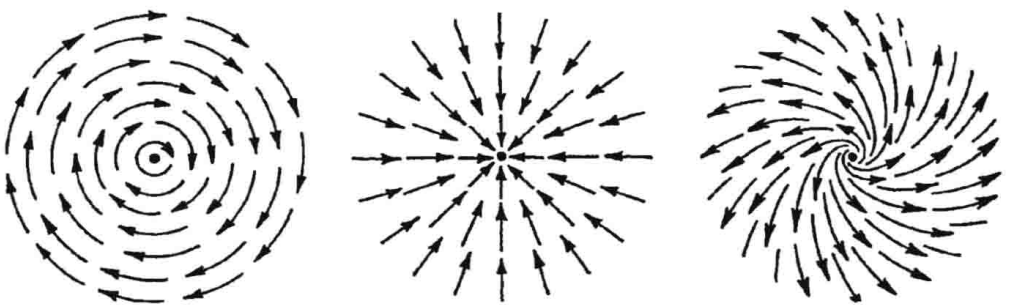


Fig. 4 Singularities in a direction field which have a count $n = +1$.

singularity whatever in the directions defined by a vector field. If the count is $+n$ (with $n > 0$) then the singularity is equivalent to n singularities like those of Fig. 4. If the count is $-n$, it is equivalent to n singularities like those of Fig. 5. If the count is zero, then by a small local distortion the singularity can be removed altogether.

The same considerations may be applied to more extended patterns. Suppose that we have a vector field defined in some region (say finite and simply-connected) in a plane. By traversing the boundary of this region once in a positive sense and (assuming no singularities lie on the boundary) counting the *net* number of complete revolutions executed by the vectors at the points of the boundary we obtain a number n (an integer – which might be negative). Thus the vectors turn by a net amount $n \times 360°$ as the boundary is traversed. This number n tells us the net count of positive singularities less the negative singularities that are encountered in the entire pattern. In Fig. 7 an illustration is given in which the count is $n = +1 = 2-1$, there being two positive singularities and one negative one. This count may be ascertained by examining the vectors only at the boundary, but, remarkably, it then gives us information about the singularity structure in the interior.

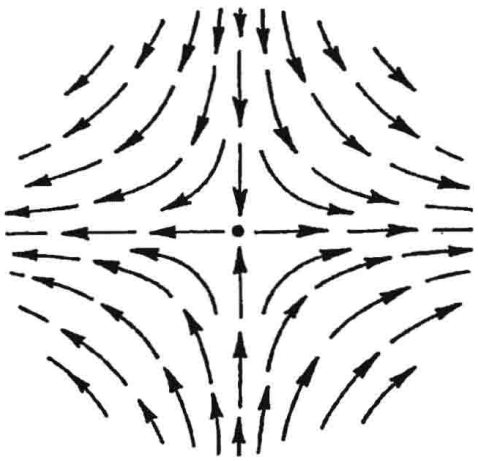


Fig. 5. A singularity in a direction field which has a count $n = -1$.

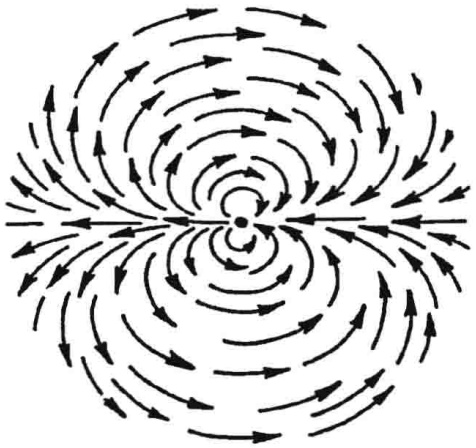


Fig. 6. A singularity with count $n = +2$.

The proof of this fact is not hard, but I shall not enter into it here. (See, for example, Lefschetz, 1962, Appendix II.) The situation may be generalized to cover vector fields which are not constrained to a plane. We may, for example, consider vector fields on the entire surface of a sphere or, say, on the surface of a ring (torus). In the first case the total singularity count gives $n = 2$ for the entire pattern and $n = 0$ in the second case. This shows, in particular, that any vector field on the sphere must contain *some* singularity in its direction field.

Let us now turn to consideration of the ridge patterns which are to be found in finger and palm prints. There is one small but important difference between such ridge patterns and the vector fields that we have just been considering, namely that whereas the elements of the vector field are *directed* the elements of ridge patterns are not. This has the effect that the basic types of singularity that can occur with ridge patterns differ in structure from those that we have encountered in the case of the vector fields. Thus we have, for these basic singularities, the loop (Fig. 8) and the triradius (Fig. 9), neither of which can occur with vector fields because the ridges cannot here be consistently orientated over the whole pattern. In fact, the presence of

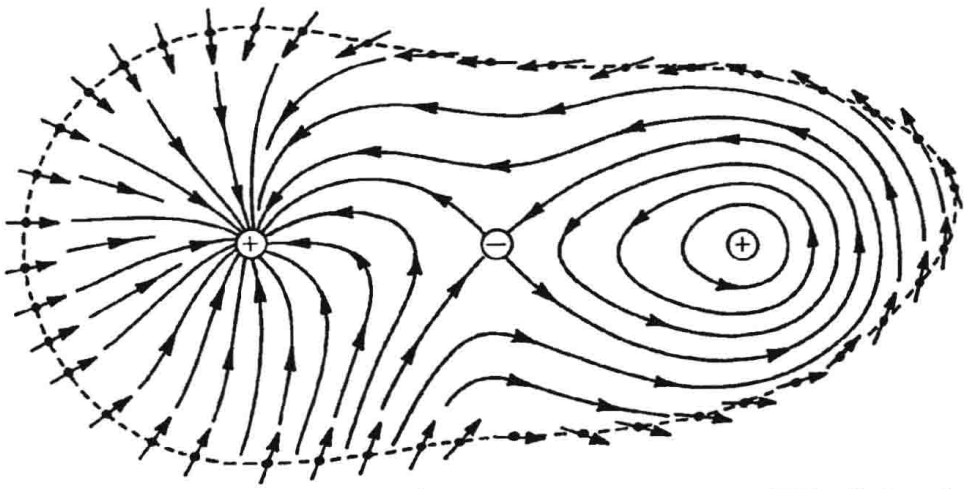


Fig. 7. The total count is $n = +1-1+1$; the arrows rotate a net amount $n \times 360°$ as the boundary is traversed.

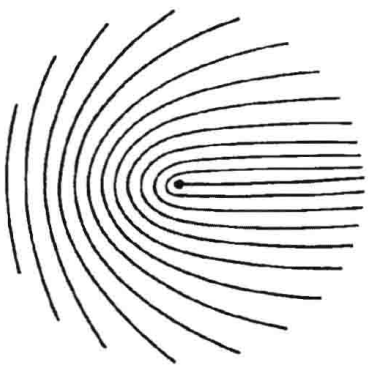


Fig. 8. A loop, having $N = +1$.

loops and triadii in fingerprint patterns may be regarded as giving us a small but significant piece of information concerning the origins of the patterns themselves (Penrose, 1967). We can infer that the ridges are not produced by a field of force, or as the lines along which some quantity takes a constant value, say. For such mechanisms would invariably produce a *directed* field where orientations could be consistently defined, and the triradius and loop would be ruled out. Instead we must expect something of a *tensor* character, such as a stress or strain, or perhaps the curvature of a surface.

The basic feature that must be present in the relevant local mechanism is that it has to be invariant under rotation through 180° (i.e. that the mechanism should possess central symmetry). This is manifestly not true of a vector field (e.g. a force), but a strain or curvature does possess such a local symmetry. The symmetry will then be present in the resulting structure of the individual ridges, thus enabling the loop and triradius configurations to be consistent.

Let us see how this affects our method of counting singularities. Applying the same ideas as before, we may imagine the singularities of Fig. 8 and 9 to be surrounded by small circles. As before, we may traverse these circles in the positive sense and observe the amount of absolute

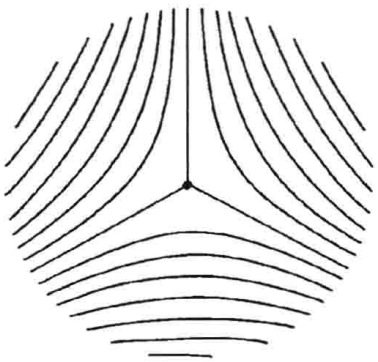


Fig. 9. A triradius, having $N = -1$.

rotation which takes place in the ridge direction when the starting point is returned to. We now find that in the case of the loop there is a rotation through 180° in the positive sense, while for the triradius the rotation of the ridge direction is 180° in the negative sense. We do not now need to obtain a complete rotation through 360° as we did for the vector fields, because of the aforementioned symmetry possessed by the ridges themselves. A rotation of a ridge through 180° returns it to a state identical with that from which it started. Thus, we may now count in terms of multiples of 180° rather than 360°. The total count as we circumnavigate a ridge pattern (on a finite simply-connected portion of a plane) will now be N where the ridges around the boundary rotate through an angle $N \times 180°$ – taking N to be an integer which can be positive, negative or zero. Thus the loop, as indicated in Fig. 8 must count as $N = +1$ and the triradius of Fig. 9 as $N = -1$.

These are the simplest types of singularity in a ridge pattern and, analogously to the case of the patterns of Fig. 4 and 5 for vector fields, we find that any localized singularity whatever in a ridge pattern can be regarded, as composed in a certain sense, of a number of loops and triradii. That is to say, by making only small local changes in the ridge pattern, and leaving the pattern unchanged at a distance from the singularity, we can produce a pattern whose only singularities are loops and triradii. Examples of ridge pattern singularities which, as they stand, are not simply loops or triradii, are the ones which would arise from the vector field patterns of Figs. 4–6 if the arrows were removed from the lines. Since with ridge patterns, we are concerned with multiples of 180° rather than 360° the counts, as ridge patterns, are now twice what they were as vector fields, that is $+2$, -2 and $+4$, respectively. To see, in one example, how such ridge pattern singularities can be slightly distorted so that only loop and triradius singularities remain, compare Fig. 10 with the final pattern of Fig. 4. Other types of singularity can be dealt with in a corresponding way.

Suppose, then, that we have a (finite, simply-connected) ridge pattern on a plane whose only singularities are L loops and T triradii. By traversing the boundary we can calculate N as above and we find

$$N = L - T.$$

This result is basically similar to the corresponding one for vector fields considered earlier. To obtain an indication of a proof, we may imagine a succession of closed curves which start as a small loop around the centre, gradually moving outwards, the final curve being the boundary

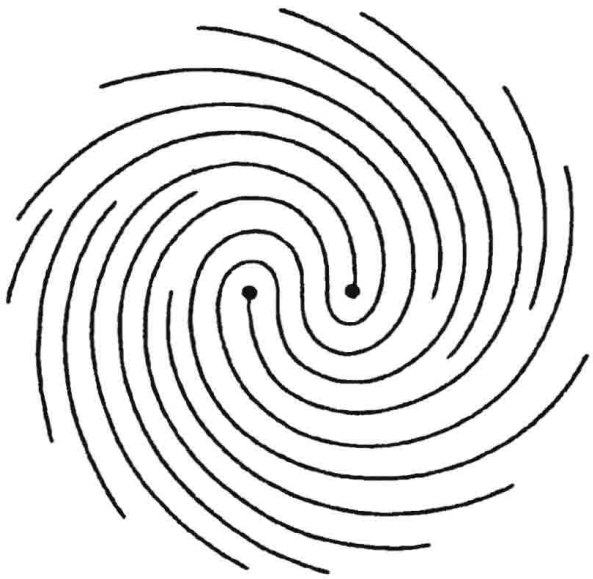


Fig. 10. Two loops giving $N = +2$, approximating the third vector field pattern of Fig. 4.

itself. For each curve c we obtain a count N_c, where $N_c \times 180°$ is the amount that the ridge direction rotates as c is traversed. Now the value of N_c, as c moves outwards, can change only when c moves across a singularity. If the singularity is a loop then N_c increases by unity; if it is a triradius, N_c decreases by unity. Finally the result $N = L-T$ is obtained when the boundary is reached.

It may be remarked in passing that corresponding results can be obtained for patterns with a different local symmetry from that of ridges. For example, we might envisage a pattern whose local structure was that of a cross, a rotation through 90° now returning the local structure to its original state. The role of the loop is taken over by the configuration drawn in Fig. 11 and that of the triradius by the one drawn in Fig. 12. The count is now taken in multiples of 90°, so Fig. 11 counts $+1$ and Fig. 12 counts -1. A result analogous to $N = L-T$ also holds for such patterns.

Let us turn to the question of ridge patterns on a normal hand. The discussion we have given does not strictly apply since hands are not flat. Nevertheless we can imagine the finger and palm prints to be stretched and flattened on to the plane. Then the earlier discussion does apply. Let us suppose that the hand (or palm) possesses D digits, that the pattern of ridges at the edges is as indicated in Fig. 13 (the normal case $D = 5$ being illustrated) and that only loop and triradius singularities occur. It will be noticed that only in the gaps between the digits does a significant rotation of the ridge direction occur, this rotation being $-180°$ at each gap. Thus, the count is

$$N = -(D-1),$$

so combining this with the earlier formula we obtain (Penrose, 1965)

$$D+L-T = 1$$

which specializes, for a normal hand, to $T-L = 4$.

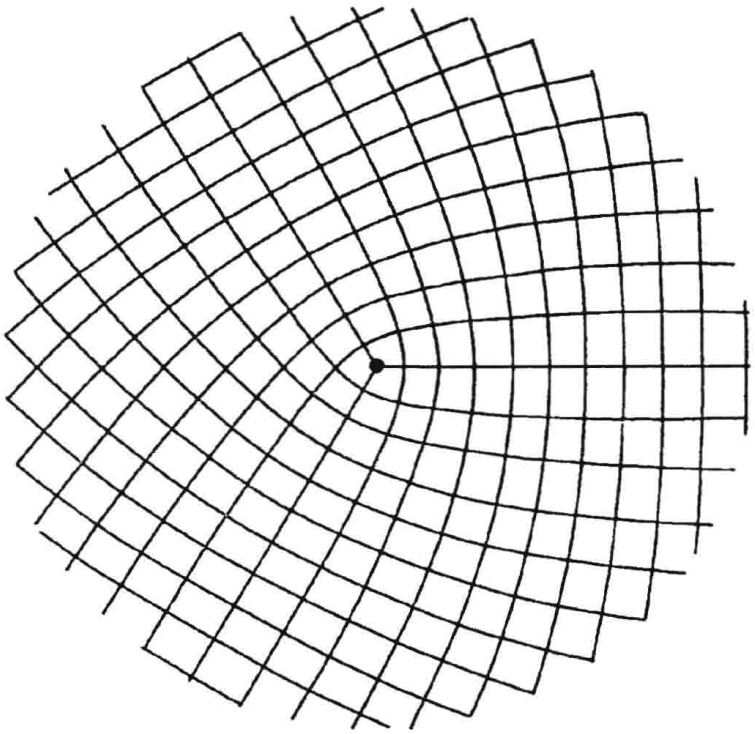


Fig. 11. A pattern built from a local structure whose symmetry is that of a cross. The singularity is at the centre and would count as +1 in this scheme.

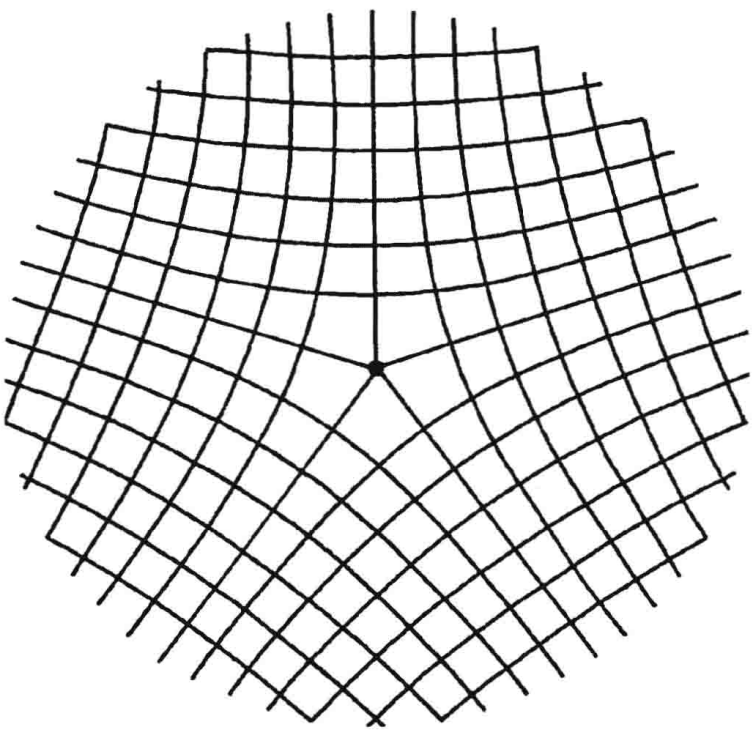


Fig. 12. A pattern like that of Fig. 11 but which would have a count −1.

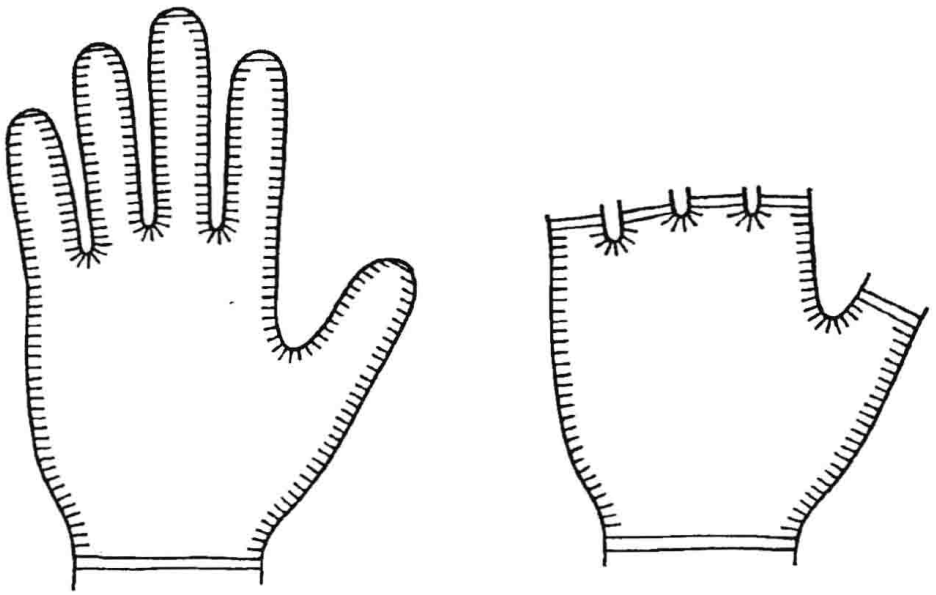


Fig. 13. The arrangement of ridge patterns around a hand or palm. The count is $N = -(D-1)$ with the number of digits being $D = 5$, here.

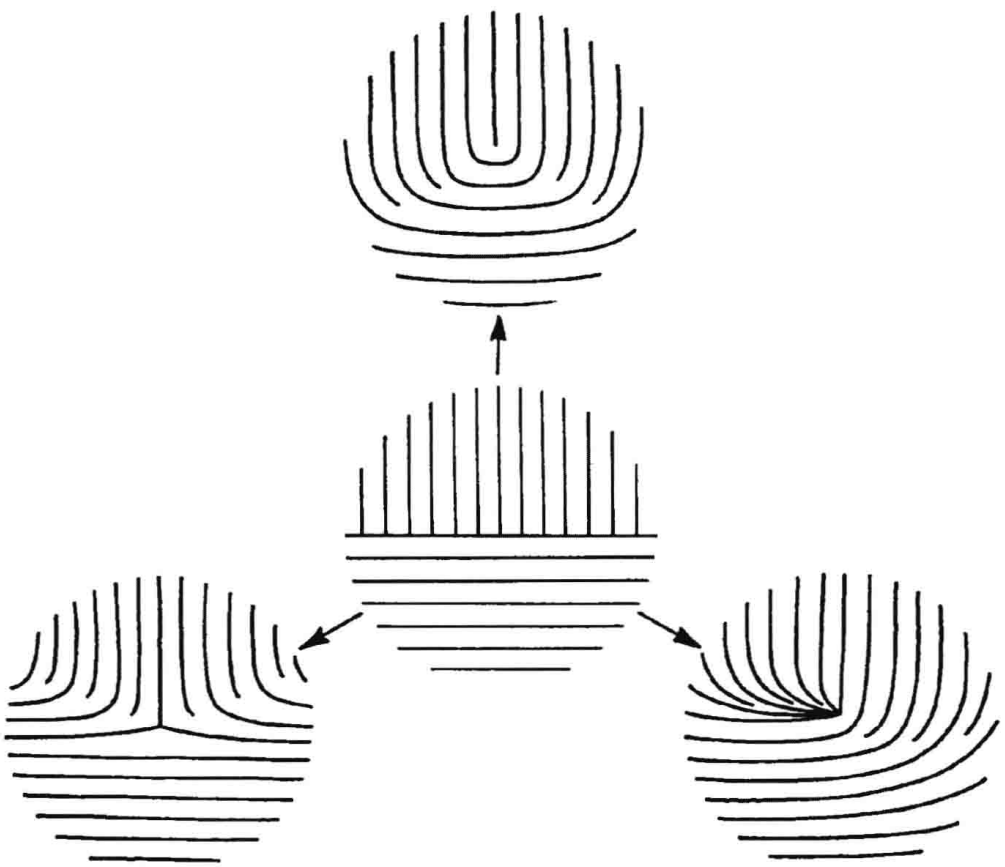


Fig. 14. Ridge patterns arising from some tensor field, such as curvature or stress, may occasionally turn rather abruptly through a right angle, as in the ridge pattern depicted at the centre. Slight variations would lead to the (more probable) other patterns shown.

Finally, I should like to touch on a matter which goes a little beyond the purely topological considerations mentioned above. If the ridge patterns arise out of something like a curvature or a stress or strain (generally, from a tensor field of the appropriate type), then we might expect to find a few places where the ridge direction abruptly changes through a right angle. This would occur when the minimum and maximum curvatures or stresses, etc. swap roles. A possible situation of this type is illustrated at the centre of Fig. 14. Varying this configuration slightly we get patterns like the others illustrated in Fig. 14. The existence of the pattern at the lower right was first pointed out to me by Smith (1979). It is topologically like a loop but rather different in its geometric structure. It would be interesting to know whether this sort of pattern occurs at all often in palm prints.

REFERENCES

LEFSCHETZ, S. (1962). *Differential Equations: Geometric Theory*. New York: Interscience, John Wiley & Sons.
PENROSE, L. S. (1965). Dermatoglyphic topology. *Nature*, **205**, 544.
PENROSE, L. S. (1967). Dermatoglyphic topology. In *Hautleisten und Krankheiten: I. Kolloq. der Deutschen Forschungsgemeinschaft* (ed. W. Hirsch). Berlin.
SMITH, C. A. B. (1979). Note on the form of dermatoglyphic patterns. In *Dermatoglyphics, 50 years later* (ed. W. Wertelecki, C. C. Plato & D. Bergsma). The American Dermatoglyphics Association and the National Foundation, March of Dimes. New York: Alan R. Liss.

This is an account, written early in 1974, of a lecture delivered in the summer of 1973 at the L. S. Penrose Memorial Colloquium in Berlin. It is dedicated to his memory. The text Figures have been prepared for publication by Mr A. J. Lee.

Mr Lee has kindly drawn my attention to the allied paper 'Fibroblast cultures and dermatoglyphics: the topology of two planar patterns' by Tom Elsdale and Frances Wasoff, published in *Wilhelm Roux's Archives of Developmental Biology*, **180**, 121–147, which applies similar topological methods both to dermatoglyphics and to fibroblasts.

101
TWISTING ROUND SPACE-TIME

New Scientist **82** (1979), *734–7*

Twisting round space-time

The paradoxical idea that matter is largely composed of empty space lies at the heart of physics. A new theory attempts to explain simultaneously both the particles and the space that make up the apparently solid world in which we live

Roger Penrose FRS
is Rouse Ball professor of mathematics at the University of Oxford

In what must be one of the most compelling introductions to a popular work of science (*The Nature of the Physical World*, based on lectures given in 1927), Sir Arthur Eddington remarked that, while relativity had caught the attention of the public with its uprooting of conventional ideas of space and time, another revolution had also taken place that he himself regarded as even more remarkable than relativity. Yet it had passed almost unnoticed by the world at large. He was referring to the discoveries which led to a description of seemingly "solid" matter, like his writing-table, as being composed almost entirely of empty space! Far removed from everyday experience as such a description might be, it accorded excellently with the scientific facts. Indeed, Eddington referred to his "two tables", one being the commonsense "solid object" of our immediate senses, and the other his "scientific table" composed ultimately of mathematical entities like the solutions of differential equations, yet infinitely more reliable for describing detailed behaviour.

In the 50 or so years since the time that Eddington wrote his book, much attention has been paid to, and much has been learned about, those minute particles of "solid" matter within atoms, which comprise no more than one part in 10^{14} of the volume of an object such as a table! (This is smaller than a grain of sand in the Royal Albert Hall). Comparatively little attention, however, has been paid to understanding the empty space of which the table is, in a certain sense, almost entirely composed. The reasons for this are perhaps, twofold, though these two reasons seem strangely opposite to one another. On one hand, there is the feeling that we completely understand the nature of empty space: it is given to great accuracy by the geometry underlying Einstein's special relativity (Minkowski geometry; *New Scientist*, 8 March, p756). On the other hand, there is the almost total lack of any really *deep* understanding of the nature of empty space.

According to the standard (quantum field) theory, empty space on a small scale is seething with activity. Every possible physical process makes a mark on it — a mark so violent that the energy density of empty space is calculated to be infinity! But this cannot actually be, particularly because Einstein's *general* relativity if directly applied to this infinite energy density would lead to an absurd picture of an infinitely tiny radius of curvature for the Universe. In quantum field theory we evade this problem by subtracting an infinite constant from the infinite energy density — a procedure known as "renormalisation" — and the finite effects remaining are supposed to be those that are observable. The hope is that some improved theory should ultimately supply a rationale for this subtraction, which is only a stop-gap procedure that reflects our ignorance.

Many years ago, Lev Landau, Wolfgang Pauli, Oscar Klein and others suggested that general relativity might, when correctly brought into line with quantum field theory, actually remove the infinities and yield a finite theory in no need of renormalisation. Then, maybe, finite observed physical quantities, such as the masses or electric charges of elementary particles, could be calculated. But it turns out that when standard techniques of quantum field theory are applied to general relativity (with matter present) the resulting theory is not even renormalisable; still less does it supply the sought-after finite theory that might serve as a quantitative replacement for the arbitrariness of renormalisation. It seems that we are as far away as ever from understanding empty space. So what prospect is there of understanding the detailed nature of the particles which inhabit that space if we do not understand the space itself?

On the scale of atoms, or their smaller constituent particles, our geometrical picture of things becomes especially inadequate. The problem is that at this level we cannot localise objects in space — a difficulty first set down by Werner Heisenberg in his *uncertainty principle*. Apparently we must relinquish geometrical pictures and rely instead on equations, if we are to retain a reliable description of reality at the subatomic level.

Equivalent descriptions

But perhaps it is not geometry itself that is at fault; perhaps it is only the specific space-time geometry that we have become accustomed to on the macroscopic scale, and which our limited imaginations suggest must also hold at the submicroscopic level, that is at fault. Yet against this we must set the successes of present-day physics. These cannot just be relinquished. If the geometrical pictures are to be changed in our descriptions of physics, then they can be changed only to something that is, in a very large measure, *equivalent* to the space-time descriptions that we are familiar with. Differences there must eventually be if such a new geometry is to represent an improvement on the old; *but it is more important to establish the measure of equivalence than the differences*, in view of the successes that physics has achieved to date.

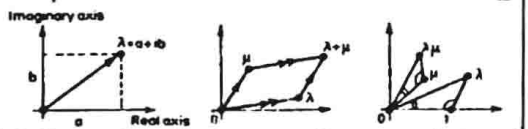
The impressive accuracy of quantum electrodynamics (the quantum field theory of electromagnetism) seems to show that our ordinary ideas of space and time must hold good down to one hundredth part of the diameter of an elementary particle (that is 10^{-17} m). Yet is this really so? We have a precedent in Newtonian gravitational theory. The accuracy of 18th-19th century celestial mechanics is perhaps more impressive even than quantum electrodynamics. Does that prove that the notion of gravitational force is a physical reality and that space

Complex numbers

A complex number has the form $a + ib$, where a and b are ordinary "real" numbers and where the symbol i stands for a quantity whose square is -1:

$$i^2 = -1$$

Complex numbers can be added, subtracted, multiplied and divided (except by zero) in the ordinary way, and they turn out to have many surprising mathematical advantages over real numbers. For the price of "inventing" just one new number, we are, for example, presented in return with systematic solutions to *all* algebraic equations. A complex number may be represented as a point on a plane known as the *Argand plane*, as shown below. This picture provides geometrical rules for addition and multiplication of complex numbers. □



Left One complex number is represented by a point on the Argand plane Centre Addition of two complex numbers. Right Multiplication of two complex numbers

is Euclidean at the dimensions relevant to planetary motions? Evidently not, for Einstein's general relativity denies both, recovers the whole of celestial mechanics and also explains some additional facts (notably the equivalence of inertial and gravitational mass and the perihelion advance of Mercury). At the level of celestial mechanics, Einstein's theory may be regarded as essentially a reformulation of Newton's. But it gives a more profound view of the nature of space, and a considerably more accurate one for the description of strong gravitational fields.

So I am suggesting that some geometric reformulation may represent a key to understanding the geometry that governs behaviour at the submicroscopic level. And in order to be able to reproduce the successful physics of our day, this reformulation must incorporate both quantum mechanics (see below) and the flat Minkowski geometry of special relativity. At a later stage, the curved geometry of general relativity must also be accommodated though in practice only a comparatively small number of physical phenomena require this.

The theory of *quantum mechanics* successfully describes the "split personality" of matter at the submicroscopic level, where waves sometimes behave like particles and particles like waves. (For example, light must sometimes be regarded as being composed of particles called "photons" in order to explain certain phenomena). The first point to recognise about quantum mechanics is the essential role played by *complex numbers* (see Box). Quantum mechanics represents physical states by objects known as *rays* in a *complex vector space*. Any two elements A, B of such a vector space can be linearly combined to form a new one

$$C = \lambda A + \mu B$$

where λ and μ are complex numbers. This is the *law of linear superposition*. Here each "vector" A, B, C (assuming that it is not the zero vector) describes the quantum state. However, the *same* state is obtained whenever the vector is multiplied by any non-zero complex number. (This is what is meant by the state being represented by a "ray" in the vector space.) From this it follows that it is only the distinct ratios of complex numbers $\lambda : \mu$, in the above combination, that provide distinct quantum states. To summarise: for any two distinct quantum states, a whole array of other allowed quantum states exists, formed from the original two by linear superposition; precisely *one* state of this array corresponds to each distinct ratio $\lambda : \mu$. This is a fundamental axiom of quantum mechanics which, to the degree that it has been possible to test it, has been fully borne out by experiment.

To understand what this array "looks like", we must fold the Argand plane, a two dimensional plot representing the family of complex numbers (see Box), into what is known as the *Riemann sphere* which includes a point "infinity". The relation between the Riemann sphere and the Argand plane is achieved by stereographic projection from the north pole of the sphere to its equatorial plane (Figure 1). Each point of the Riemann sphere represents a complex ratio, ie $\kappa = \lambda/\mu$, the north pole representing an infinite value for κ corresponding to $\mu = 0$, and every other point of the sphere projects down to a unique point $\kappa = a + ib$ of the (equatorial) Argand plane. So, the north pole represents the quantum state A, the south pole represents the state B, and the point labelled κ represents the state $C = \lambda A + \mu B$ (Figure 2). In this way linear superposition of any two quantum states gives rise to an array of states which has the topological structure of a sphere.

Spinning particles

Though this is usually just an abstract mathematical sphere, a relation between this sphere and spatial geometry does exist, and can be understood in terms of the physical states of a spinning particle, such as an electron or proton. If in a state A the particle spins right-handedly about the upward vertical, and in state B about the downward vertical, then in state C the particle spins right-handedly about the direction represented by $\kappa = \lambda/\mu$. In other words the spin-directions in space

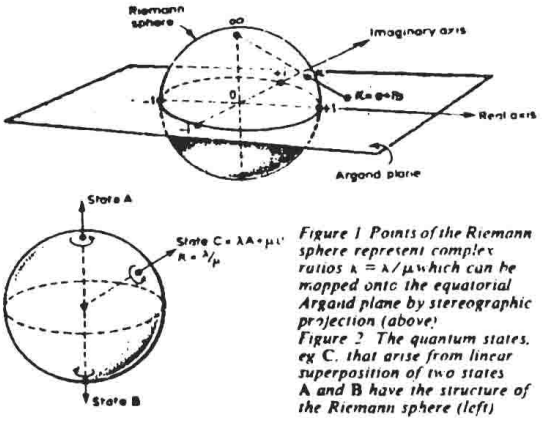


Figure 1 Points of the Riemann sphere represent complex ratios $\lambda = \lambda/\mu$ which can be mapped onto the equatorial Argand plane by stereographic projection (above)
Figure 2 The quantum states, eg C, that arise from linear superposition of two states A and B have the structure of the Riemann sphere (left)

correspond exactly to the points of Riemann sphere, through linear superposition of the states of the simplest type of spinning system. So the *three-dimensionality of space* is related intimately to the fact that *complex numbers are used in the linear superposition law.*

The simplest way to see how space-time geometry is also intimately related to complex structure, is to re-interpret the Riemann sphere as the celestial sphere or "sky" of some observer who, for the purpose of our descriptions, we imagine to be out in space. He thinks of himself as positioned at the centre of the Riemann sphere, and he labels each star that he sees in his sky by the complex ratio, κ, corresponding to the direction that he sees the light to be coming from. There is an arbitrariness in how he orients his Riemann sphere against the sky. There is also the "relativistic abberation" effect: this implies that the geometry of the apparent positions of the stars that he sees actually depends on his velocity at that moment. But the transformation from one choice of orientation and velocity giving a ratio κ, to another giving κ' is provided by a remarkably simple formula:

$$\kappa' = \frac{\alpha\ \kappa + \beta}{\gamma\ \kappa + \delta}$$

Here α, β, γ and δ are fixed complex numbers defined by the relative orientations and velocities between the two choices.

An important feature of this formula is that it is what is called *holomorphic*. In simple terms "holomorphic" means that the formula involves only terms of the form $\kappa = a + ib$, the "complex conjugate" $\bar{\kappa} = a - ib$ being absent from the expression. Holomorphicity, which means, more generally, "complex smoothness," is a key concept in the theory that will follow, as indeed it is in much of quantum field theory and particle physics. In fact one aim of twistor theory is actually to reduce all the equations of physics simply to "pure holomorphicity". An example of this is provided by the above transformation for κ which turns out actually to be the most general purely holomorphic transformation sending the Riemann sphere to itself. This group of transformations *(New Scientist,* 5 April, p34) is the "restricted Lorentz group", which is the most fundamental symmetry group in physics, describing the special relativistic symmetry of space-time about a point.

We now relate this to space-time geometry. The observer, at the instant that he observes his sky, can be represented as a point *O* in four-dimensional (Minkowski) space-time; the light rays that enter his eye are represented by the "past light cone", as shown in Figure 3. In this kind of space-time diagram, time corresponds to the direction "up" the page and the dimensions of space are represented by directions at right angles to this. At any point in space time such as the point *O*, the system of light rays reaching an observer there sweeps out ("generates") a cone stretching back into the past. Each one of these light rays

corresponds to one point of the celestial sphere — which we think of as the Riemann sphere of κ. But imagine a new space (PT in what follows) each single point of which represents an entire light ray in Minkowski space-time. The light rays through O will then be represented by a subset of points in this new space. Now, the Riemann sphere is a *complex* space (of one complex dimension) — is the *entire* light-ray space also a complex space?

Not quite. Any complex space must be of an even number of dimensions, because each complex dimension is made up of two real dimensions. Both the Argand plane and Riemann sphere, for example, are made up of *two* dimensions of real space, although they are defined by a *single* complex parameter. But light rays actually form a 5-dimensional system location and direction). To achieve a new "even-dimensional" complex space we have to replace our light rays by spinning photons, the "particles" of light, which also possess energy. They can spin only by a fixed amount in a right-handed or left-handed manner, so the extra variable quantity is the energy. This gives an abstract 6-dimensional space whose points represent spinning photons. And very remarkably, it turns out that this space can indeed be regarded as a complex 3-dimensional space in terms of which the symmetry transformations of space-time are purely holomorphic.

This space, PT is called *projective twistor space*. It is a higher-dimensional version of the Riemann sphere, its points again being given in terms of ratios, but this time they are ratios of *four* complex numbers, namely $Z^0:Z^1:Z^2:Z^3$, which are the components of what we call a *twistor* represented as Z^α. Projective twistor space divides into three regions, labelled PT^+, PT^- and PN (Figure 4).

The points of PT^+ represent photons spinning in a right-handed sense about their directions of motion, and the points of PT^-, those spinning in a left-hand sense. PN represents the real five-dimensional boundary between the two halves PT^+ and PT^-; its points give the limiting case when the energy is taken to be infinite and the relatively small effects of spin can be neglected. So the points of PN in fact represent "light rays". But when spin is present, the localised description of a photon in a light-ray is just an approximation. The spinning photon acquires a "spread-out" (non-local) structure, determined by its motion, that can be described geometrically in space-time terms as a system of twisting lines known as a "Robinson congruence". The spatial description in the right-hand case is indicated in Figure 5, where the whole picture moves upwards with the speed of light.

Though I have given a somewhat "physical" description of the space PT in terms of spinning photons (and, indeed,

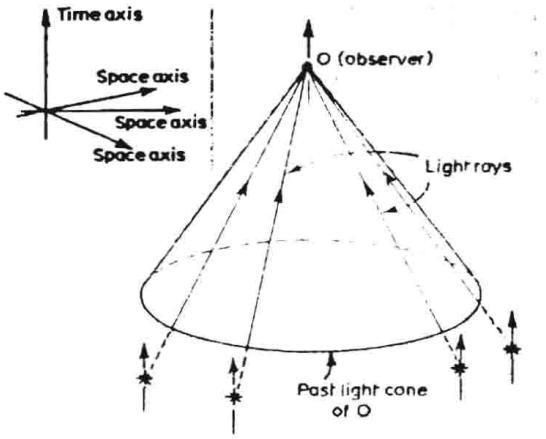


Figure 3 Minkowski space-time, depicting the past light cone *of an observer at* O *(above)*

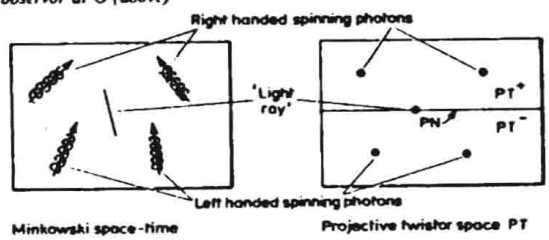


Figure 4 In projective twistor space points *represent spinning photons; these are particle histories in Minkowski space-time*

physical photons do possess a spin with just these properties), this is actually a slightly misleading way of regarding a twistor. A photon is just one among many different families of known particles. Twistors do not have any special relation to any particular kind of particle, but are something more basic than the concept of a particle. More correctly, they simply provide an alternative, though more "particle-like", way of viewing the geometry of space-time. The geometry of twistor space provides a new basis from which space-time can be derived.

So we work with PT and dispense with the space-time

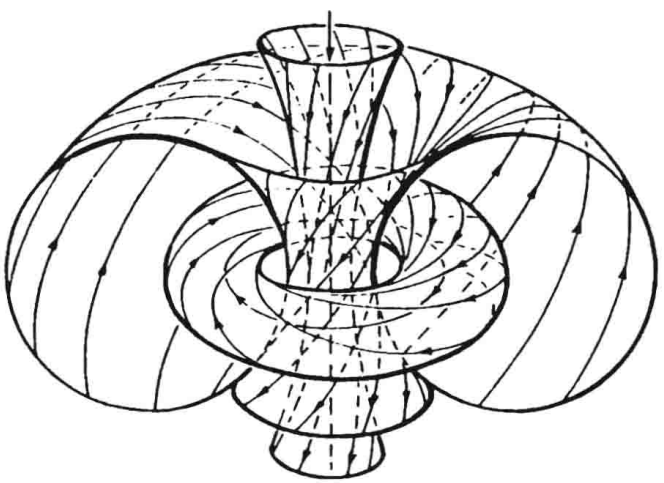


Figure 5 In space-time the kinematics of a spinning mass-less particle, such as a photon, is described by a complicated system of lines called a Robinson congruence. *This diagram, shows the spatial description of such a system which moves up the page with the velocity of light*

picture. In fact, the whole of Minkowski geometry could be reconstructed again from the twistor space (the space-time events being reinterpreted as certain "lines" lying in the space PN). But the essential aim of twistor theory is to try to describe as much as possible in purely *twistor* terms. The idea is that the more fundamental the level at which physics is being treated, the more relevant will be the descriptions in terms of twistors and the less relevant will be those given in terms of space-time geometry.

A more detailed aim of twistor theory is to transcribe conventional quantum field theory into twistor terms with the hope of obtaining a finite theory, free of the infinities that plague the conventional approach. There are many mathematical difficulties that still stand in the way of this programme, though it is possible to see ways of maintaining finiteness, if certain rules can be followed. As part of the programme, a twistor theory is being used to describe elementary particles. Massive particles, such as electrons or protons can be handled as well as massless ones such as photons. But while a single-twistor description is adequate for a massless particle, two or more twistors are needed for a massive one.

One of the striking features of these descriptions of particles is that certain "internal symmetry groups" arise in a quite natural way in the theory. When treated quantum mechanically, these groups provide "internal quantum numbers" that can be used to classify the different kinds of particles. In the most obvious scheme leptons (eg electrons, muons, and their neutrinos, *New Scientist*, 22 February, p564) are described in terms of two twistors while hadrons (eg protons, neutrons and pions, *New Scientist*, 1 March, p664) require three or more. This gives the leptons a "pointlike" character, whereas hadrons would appear as more extended objects. The scheme seems to be in reasonable accord with what is known from particle physics, but is not yet so sufficiently well-defined that clear-cut predictions can be made. The main work concerning this programme is being carried out by George Sparling, Zoltan Perjes, Lane Hughston, and Andrew Hodges in conjunction with myself and several others.

The twistor description of *general* relativity leads to some striking results, though this work is far from complete. Roughly speaking, the curving of Minkowski space-time corresponds to deforming the twistor space. There is a sense in which the curvature of empty space-time can be split into two pieces, called the self-dual part and the anti-self-dual part. It turns out that the anti-self-dual part can be completely understood in terms of these deformations, Einstein's equations being, in a sense, reduced to pure holomorphicity in accordance with the general programme. It now seems quite possible that the self-dual part can also be understood in terms of a different kind of twistor space deformation, but the complete analysis of this requires a considerable amount of further work.

It appears likely that, in a similar way, the other interactions of nature can be understood in terms of twistor space deformations. Indeed, Richard Ward has shown how to do this, in the anti-self dual case, for electromagnetism and for the "Yang-Mills" fields that are believed to be responsible for elementary particle interactions.

Twistor theory is incomplete, but it offers considerable hope for a quite new approach to the basic problems of theoretical physics. It seeks to explain both the minute particles and the vast empty space that are together needed to compose Eddington's "scientific writing-table". The twistors are to be more basic than either; the correct dimensionality of space and time, for example, being at least as much a necessary consequence of the twistor approach as is any particular property of particles. Twistors are abstract mathematical entities whose properties are forced by abstract mathematical principles, such as holomorphicity with its rich and subtle geometrical consequences. If such a view of the ultimate nature of physical reality is correct, then we, like Eddington's scientific table, are all composed of abstract mathematics! □

102
A GOOGLY GRAVITON?

Twistor Newsletter **8** (1979). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

This seems to have been my first serious attempt to tackle the gravitational googly problem (see Chapters 70 and 82). Figure B on p. [173] was sometimes referred to by my colleagues as 'the rotten banana at infinity'.

 A Googly Graviton? *Twistor Newsletter* **8** (1979). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

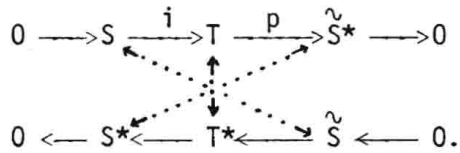
§3.9. A GOOGLY GRAVITON? *by R. Penrose*

The so-called "non-linear graviton construction" (cf. §3.1c) enables the local structure of any anti-self-dual complex solution of Einstein's vacuum equations to be encoded into the global structure of a deformed twistor space. The weak field limit of this construction yields the $H^1(..., \mathcal{O}(2))$ "twistor function" description of linearized gravity (helicity -2). For weak fields we also have the $H^1(..., \mathcal{O}(-6))$ "twistor function" description in the *self*-dual case (helicity +2). Is there a non-linear version of this, providing a deformed-twistor-space method of describing *self*-dual complex Einstein vacuums? (Of course, one could apply the original non-linear graviton construction to *dual* twistor space, but that is not what is meant here.) If so, then there would be the hope that by combining both types of deformation, one might be able to encode the local (?) structure of a *general* solution of the vacuum equations (i.e. just $R_{ab} = 0$)!

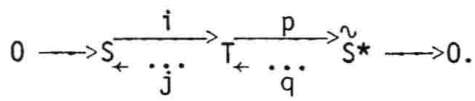
But there would seem to be an insurmountable problem right at the start. The key to the non-linear graviton construction lies in the interpretation of a point in the complex "space-time" M as a compact holomorphic curve in PT (T being a suitable deformed portion of flat twistor space T), with null separation in M being defined as intersection of the corresponding holomorphic curves in PT. This leads at once to the fact that M is conformally anti-self-dual, because fixing a point in PT we get a 2-parameter family of curves through it, providing a corresponding α-surface in M — 3 parameters worth of α-surfaces in all. Altering the definition of "null separation" seems unpromising and to redefine these as β-surfaces would be cheating. However, an alteration in the definition of a "space-time point" to obtain the opposite helicity is suggested by the following.

Consider the standard exact sequence (Penrose & MacCallum 1972) which expresses the Poincaré invariant structure of *flat* twistor space T, where S is spin-space, * means dual, $\sim$ means complex conjugation, $\leftarrow\cdots\cdots\rightarrow$ means isomorphic via complex conjugation, and where the maps i and p are defined by $i : \omega^A \mapsto (\omega^A, 0)$, $p : (\omega^A, \pi_{A'}) \mapsto \pi_{A'}$

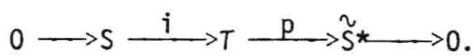
and *dually*



Now, the non-linear graviton construction for anti-self-dual vacuums depends upon the existence of a deformed version of the map $p : T \to \tilde{S}^*$, this being the given holomorphic fibration. The points of M are *cross-sections* of this fibration (to which the Euler vector field T is tangent). In the flat case, we can think of these as linear maps q back from $\tilde{S}^*$ to T such that the composition $p \circ q$ is the identity of $\tilde{S}^*$ and which provide one way of "splitting" the sequence, i.e. of expressing T as a product, compatibly with the sequence



Quite equivalent, in this flat case, is to "split" the sequence in the dual way with a map j from T to S such that $j \circ i$ is the identity on S. We may envisage a "deformed" version of this situation, however, in which T loses its linear space structure and is replaced by some suitable complex manifold $\mathcal{T}$ which, in some suitable sense, has structure providing an exact sequence



We know, from the original non-linear graviton, that the definition of

"space-time point" analogous to the q-maps will provide us with an anti-self-dual M. Furthermore, because complex conjugation and duality are so intimately related in twistor theory, we may anticipate that an appropriate analogue of the j-maps will provide us with a space $\tilde{M}$ that is a (general) self-dual complex vacuum.

To see what this is likely to entail, consider the flat case again and, for simplicity of description, we examine the projective space PT. (Fig. A)

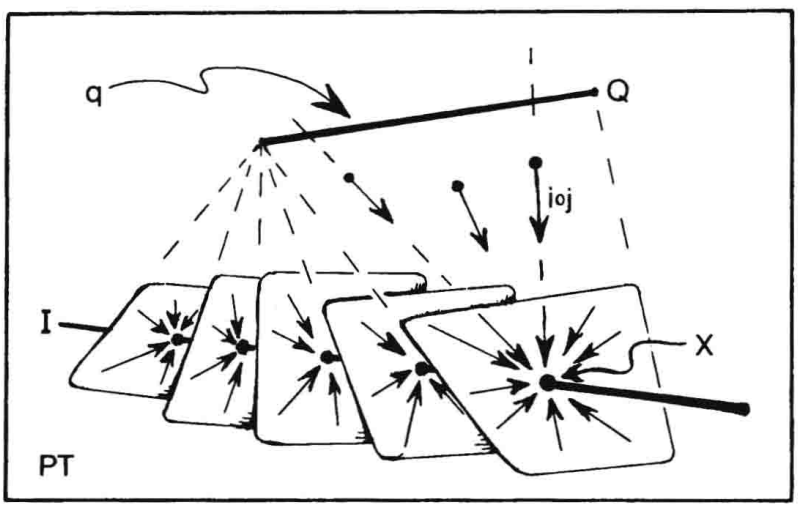


Figure A

The various maps q correspond to the various lines in PT not meeting the line I. Now I corresponds to the image of i, so each map j (or, more correctly, $i \circ j$) may be thought of as a projection from PT down onto I. The inverse image of the zero of S ($j^{-1}(0)$) is, however, not represented in this projection. This is the line Q in PT that arises in the above

"q-map". Every other point of PT - Q takes part in the "i o j" projection. It is mapped to where the plane joining it to Q meets the line I. Thus, the inverse image of any point X of I under this map is a *plane* through Q and through the point X (excluding the line Q itself). This family of planes, each containing a different point X on I and none containing I itself, provides another way of representing the line Q, and hence another way of representing a space-time point. It is essentially the *dual* description to that of the point-locus Q since here we have a "plane-locus", namely a point-locus in PT*. But when PT is deformed to the space PT, we do not expect planes to survive as global loci. Indeed, it is this fact that enables the "dual" construction to differ from the original non-linear graviton one. The "planes" need only exist in the neighbourhood of I and do not need to extend far enough out that they actually have a holomorphic curve "Q" in common. At each point X of I there will be a two-parameter family of possible "planes" through X. The totality of these "planes" constitutes a complex 3-manifold which is a fibration over I - giving the space PT* - and the "dual points", giving the space $\tilde{M}$, would be the cross-sections of this fibration PT*. For this "dual" construction to work, therefore, we require that the space PT* that is to arise in this way be similar in structure to the usual space "PT" of the original non-linear graviton. One important feature of this structure, however, is that the fibres are *not* flat spaces. If they were, then we should have, in effect, a vector bundle over CP^1 and all information would be lost, the cross-sections providing a manifold with only a flat structure. But if our deformed space PT is too "well-behaved" near I, then this is exactly what will happen. Our "planes", being well-defined near I, would be expected to have good tangent planes at their intersections with I, and to be

determined by these tangent planes. Regularity at I would imply that the required PT* space is simply the canonical *flat* one and all our supposed information concerning self-dual curvature would be lost. Thus, for this "dual" construction to work, it is necessary that the neighbourhood of I in PT be in some sense "singular" or "peculiar".

In order to investigate what kind of peculiarity to expect, consider a space-time W that is about as well-behaved as one could possibly ask for. W is to be real, (+ - - -), analytic, vacuum, asymptotically simple (Penrose 1965), with a regular point i^+ (future timelike infinity) and a point i^0 (spacelike infinity) that is as regular as i^+ when viewed intrinsically from within $\mathscr{J}^+$ (though i^0 would be singular for W). It seems that many such space-times *ought* to exist, having the full functional freedom of two real functions of three real variables. They describe imploding-exploding free gravitational waves (and, of course, Minkowski space is a special case). The curved twistor space *T* that we are trying to obtain is to be the *asymptotic twistor space of W*. Then the original non-linear graviton construction will yield Newman's $\tilde{H}$-space as our M. The idea is that, in some appropriate sense, the above "dual" construction ought to yield Newman's *H*-space as our $\tilde{M}$. In order for this to work, the information of the structure of *dual* twistor space must also be contained in *T* and by the preceding discussion, it would be expected to reside in the peculiar way that the line I sits in its immediate neighbourhood in *T*.

Recall that the points of PT-I are represented on $C\mathscr{J}^+$ by "twistor lines", that is, by null geodesics, other than generators of $C\mathscr{J}^+$, that lie in the β-planes on $C\mathscr{J}^+$ (Penrose & MacCallum 1972, Penrose 1975). (Fig. B). Each point of I would be represented by an entire α-plane on $\mathscr{J}^+$. For each fixed β-plane on $C\mathscr{J}^+$, the twistor lines lying in it form a 2-parameter

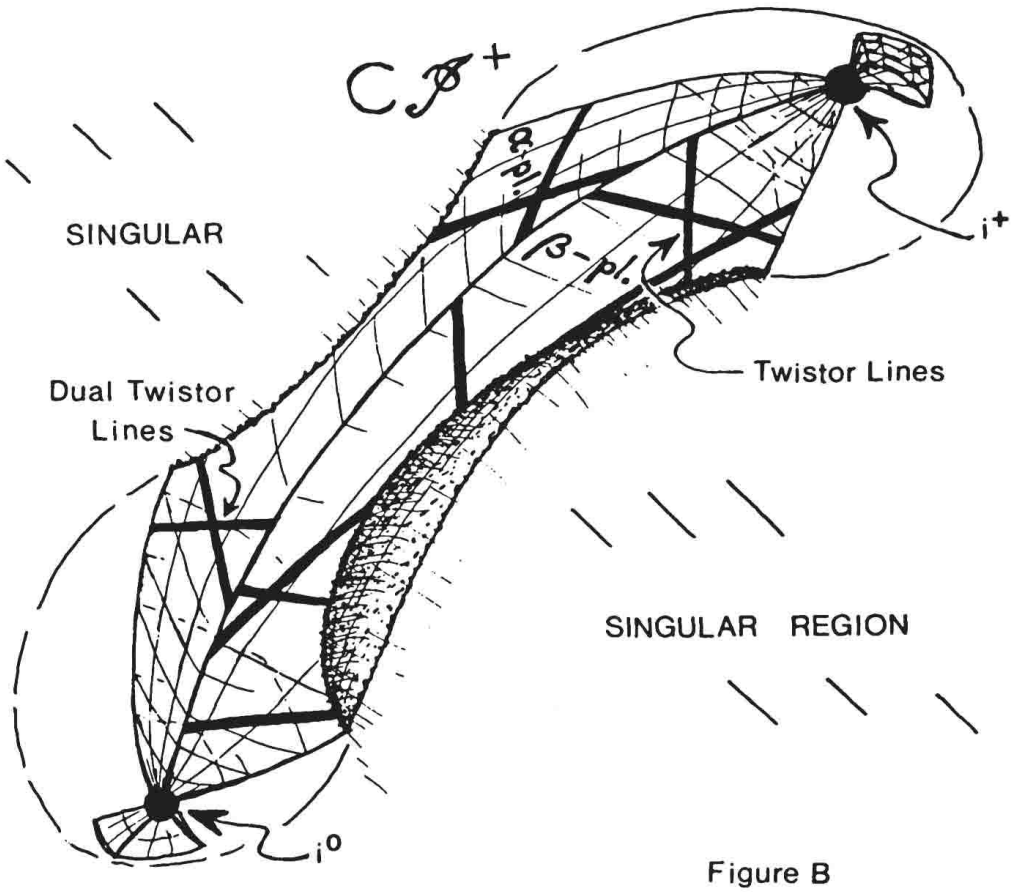


Figure B

family representing the points of a special 2-surface in PT through I. Now consider a sequence of points, on such a special 2-surface, which approach a point X of I. This is represented by a sequence of twistor lines, in the corresponding fixed β-plane, which turn around and approach a generator of $C\mathcal{J}^+$, this generator being the intersection of the α-plane ξ on $C\mathcal{J}^+$, representing X, with the given β-plane. If we vary the special 2-surface through I, the β-plane on $C\mathcal{J}^+$ correspondingly varies, so we get different generators of ξ as the limit of the sequence of twistor lines. It is thus by no means obvious that the points of I need be regular points of PT.

We have assumed, however, that i^+ is a regular point of $C\mathscr{J}^+$. It is not unreasonable, therefore, to suppose that the family of twistor lines that pass through the immediate vicinity of i^+ will in fact, give rise to a region of PT - call it PT_+ - in which the line I lies imbedded in a regular way. But we have also assumed that i^0 is a regular point of $C\mathscr{J}^+$. The twistor lines that pass through the immediate vicinity of i^0 should, therefore, also provide a region of PT in which I lies regularly imbedded - call this one PT_-. However there will be twistor lines passing arbitrarily close to i^0 that do not reach the vicinity i^+ and *vice versa*. This is because, roughly speaking, they are disconnected by the "singular region" of $C\mathscr{J}^+$. On the other hand, there are other twistor lines, namely those close to *real* generators of $C\mathscr{J}^+$, which pass arbitrarily close both to i^+ and to i^0. As such lines vary, there comes a point at which they become disconnected. (Fig. B). Thus we have a typical situation giving rise to a *non-Hausdorff* manifold PT, having regions in which bifurcation takes place. In fact, in order that PT be a manifold at all, it is necessary that there should be *two copies* of the line I, labelled I_+ and I_0, where I_+ sits regularly imbedded in PT_+ and I_0 regularly imbedded in PT_0. The open manifolds PT_+ and PT_0 can each be Hausdorff, but they are glued together (because of the existence of twistor lines passing close *both* to i^+ and to i^0) in a way that results in a non-Hausdorff combined manifold.

We may ask whether this non-Hausdorff feature is *essential* or not, i.e. whether the gluing of PT_+ to PT_0 cannot or can be extended until the resulting manifold is Hausdorff. The answer is that it *is* essential (at least in the general case under consideration). To see this, we recall a statement made earlier to the effect that a *regular* imbedding of I in PT would lead to a (conformally) *flat* structure for $\tilde{M}$. As applied to PT_+,

this would give us the complex Minkowski space $\tilde{M}_+$ that arises from the "good cuts" of $C\mathscr{J}^+$ arbitrarily near i^+; as applied to PT_0 this gives another complex Minkowski space $\tilde{M}_0$ arising from "good cuts" arbitrarily near i^0 (cf. Newman & Penrose 1966). When W is non-flat we expect that the presence of outgoing gravitational radiation in W will imply that the good cuts near i^+ and those near i^0 will be related by a non-trivial *supertranslation* (Sachs 1962) at least in the general case, but perhaps always. The complex Minkowski spaces $\tilde{M}_+$ and $\tilde{M}_0$, which describe, in effect, the first neighbourhoods of I_+ in PT_+ and of I_0 in PT_0, respectively, thus differ from one another in an essential way and cannot be identified. It follows that I_+ cannot be identified with I_0 while retaining a smooth complex manifold structure, so PT seems to be an *essentially* non-Hausdorff manifold.

The presence of the above supertranslation tells us that the first neighbourhoods of I_+ and I_0 differ. The structure of this difference is described in terms of a holomorphic function of two variables (e.g. of the usual ζ and $\tilde{\zeta}$ coordinates for $C\mathscr{J}^+$). This function is obtainable, by integration, from the required self-dual radiation field Ψ_4^0 (Newman & Penrose 1962) which will determine the structure of $\tilde{M}$, but represents only *part* of the information contained in Ψ_4^0 - necessarily, because Ψ_4^0 is a function of *three* variables instead of two. But the non-Hausdorff feature of PT involves an incompatibility not only between the first neighbourhood of I_+ and I_0 but between their second and higher neighbourhood as well. Taking all this into account, it seems possible that the detailed non-Hausdorff nature of PT does indeed contain sufficient information that Ψ_4^0, and thus $\tilde{M}$, can, in principle, be determined.

Two immediate problems for this programme suggest themselves. In the first place, we may examine "infinitesimally non-Hausdorff spaces" (i.e.

where the map describing the gluing of T_+ to T_0 is given by an infinitesimal displacement of a portion Q of flat twistor space T), to see whether these correspond to cohomology groups $H^1(PQ, \mathcal{O}(-6))$, with $PN \subset PQ \subset PT$, in any way. In the second, we must find a way to characterize the required "j-maps" so that they indeed provide the sought-for space $\tilde{M}$. This involves finding a suitable regularity condition for these j-maps that contains the information of the location of the *dual* twistor lines (lying on α-planes on $C\mathscr{I}^+$). The idea is that this information ought already to be contained in the non-Hausdorff nature of PT (or perhaps of T; the j-maps are presumably to be defined on the space $T_+ \cap T_0$). More generally, a thorough study of non-Hausdorff (but well-behaved) complex manifolds seems called for. A further problem is to find a *third* definition of "space-time point" (perhaps some kind of "average" of j and q) to yield W itself - a generic solution of Einstein's vacuum equations!

REFERENCES

Newman, E.T. & Penrose, R. 1962 J. Math. Phys. 3, 566.

Newman, E.T. & Penrose, R. 1966 J. Math. Phys. 7, 863-870.

Penrose, R. 1965 Proc. Roy. Soc. A284, 159.

Penrose, R. 1975 *Twistor Theory, its Aims and Achievements.* In: *Quantum Gravity,* eds. C.J. Isham, R. Penrose & D.W. Sciama. Oxford: University Press.

Penrose, R. & MacCallum, M.A.H. 1972 Phys. Reports 6C, 241-315.

Sachs, R.K. 1962 Phys. Rev. 128, 2851.

103
TWISTORS AS HELICITY RAISING OPERATORS

Twistor Newsletter **8** (1979). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

This gives a somewhat different perspective on the physical interpretation of a twistor. It has the significance that the complex-linearity of twistor space is directly seen to be the *same* complex-linearity as that of quantum mechanics, a fact that is by no means obvious in other interpretations of a twistor. For example, in the introductory section of Chapter 69, and elsewhere, the complex structure (Riemann sphere) of the space of null directions at a point and the complex quantum superpositions of spin states of a spin-½ particle, one classical and the other quantum, are related to one another, but it is not clear that these could be fairly regarded as being the 'same' Riemann sphere, especially since one is a conformal sphere and the other a metric one.

 Twistors as Helicity Raising Operators. *Twistor Newsletter* **8** (1979). Also published in *Advances in Twistor Theory.* Ed. L. P. Hughston, Pitman, 1979

§2.13 TWISTORS AS HELICITY RAISING OPERATORS *by R. Penrose*

Twistors have a bewildering number of different-looking interpretations, e.g. Robinson congruences, spinning massless particles (kinematics), α-planes in complexified space-time,... . In most of these there is no evident connection between the complex linearity of twistor space and quantum complex linear superposition (and, indeed, in most, the twistor phase is not even represented), whereas it has been one of the primary stated motivations behind twistor theory that twistors should provide an essential link between the complex linearity of quantum theory and the geometry of space-time. There is, however, one interpretation of a twistor, namely as a *helicity raising operator*, where the relation between twistor linearity and quantum linear superposition becomes quite apparent. The same remark applies to the interpretation of a dual twistor as a *helicity lowering operator*.

Let $f_{ij} = f_{ij}(Z^\alpha) = -f_{ji}$, $\rho_{[i}f_{jk]} = 0$ be a Čech cocycle for some covering $\{U_i\}$ of T^+, so that if f_{ij} has homogeneity $-n-2$ it is a twistor wave function (not necessarily normalizable) for a free massless particle of helicity $\frac{n}{2}$. Now consider a twistor P^α and a dual twistor Q_α. These provide new twistor wave functions $(P^\alpha \partial/\partial Z^\alpha)\, f_{ij}$ and $(Q_\alpha Z^\alpha) f_{ij}$ for particles of helicity $\frac{n}{2} + \frac{1}{2}$ and $\frac{n}{2} - \frac{1}{2}$, respectively. (The operations obviously apply to the cohomology classes and not just to the representatives.) Quantum linear superposition refers to the linear structure of the f's (i.e. states f_{ij} and g_{ij} superpose linearly to give $\lambda f_{ij} + \mu g_{ij}$; $\lambda, \mu \in C$), and this clearly relates to the linearity in P^α and Q_α. If the (complex) space-time field description of f_{ij} is $\psi_{\underbrace{A'B'\ldots}_{n}L'}$ $(n \geq 0)$ or $\phi_{\underbrace{AB\ldots}_{-n}L}$ $(n \leq 0)$, then that of $(P^\alpha \partial/\partial Z^\alpha) f_{ij}$ is

$$-P^M\nabla_{MM'}\,\psi_{A'B'\ldots L'}-(n+1)P_{(M'}\psi_{A'B'\ldots L')}(n\geq 0) \text{ or } -P^L\phi_{AB\ldots KL}(n\leq -1),$$

respectively, and that of $(Q_\alpha Z^\alpha)f_{ij}$ is

$$Q^{L'}\psi_{A'B'\ldots K'L'}(n\geq 1) \text{ or } -iQ^{M'}\nabla_{MM'} - (n-1)Q_{(M}\psi_{AB\ldots L)}(n\leq 0),$$

respectively, where the pair of spinor fields $(P^A,P_{A'})$ represents P^α and the pair of spinor fields $(Q_A,Q^{A'})$ represents Q_α, with $\nabla_{AA'}P^B = -i\varepsilon_A{}^B P_{A'}$, $\nabla_{AA'}Q^{B'} = i\varepsilon_{A'}{}^{B'}Q_A$ (see Penrose 1975 and §2.14).

We investigate the kernel and quotient of these maps

$$\Phi_n \xrightarrow{P^\alpha} \Phi_{n+1}\,, \qquad \Phi_n \xrightarrow{Q_\alpha} \Phi_{n-1}\,,$$

where Φ_n stands for the space of helicity $\frac{n}{2}$ massless wave functions (not necessarily normalizable), i.e. equivalently, $\Phi_n = H^1(PT^+, \mathcal{O}(-n-2))$. We use the short exact sequences of sheaves over PT:

$$0 \to \mathcal{O}^P(-n-2) \to \mathcal{O}(-n-2) \xrightarrow{P^\alpha\partial/\partial Z^\alpha} \mathcal{O}(-n-3) \to 0,$$
$$0 \to \mathcal{O}(-n-2) \xrightarrow{Q_\alpha Z^\alpha} \mathcal{O}(-n-1) \to \mathcal{O}_Q(-n-1) \to 0,$$

where $\mathcal{O}^P(-n-2)$ refers to homogeneous degree $-n-2$ functions which are constant along the linear 2-spaces in T through P^α, and where $\mathcal{O}_Q(-n-1)$ refers to homogeneous degree $-n-1$ functions on the 3-space $Q_\alpha Z^\alpha = 0$, so in each case these describe twisted holomorphic functions on a CP^2. The corresponding long exact sequences are (where in each case the space under consideration is PT^+)

$$\ldots \to H^1(\mathcal{O}^P(-n-2)) \to \Phi_n \xrightarrow{P^\alpha} \Phi_{n+1} \to H^2(\mathcal{O}^P(-n-2)) \to \ldots$$

and

$$0 \to H^0(\mathcal{O}(-n-2)) \to H\,(\mathcal{O}(-n-1)) \to H^0(\mathcal{O}_Q(-n-1)) \to \Phi_n \xrightarrow{Q_\alpha} \Phi_{n-1} \to$$

$$\to \quad H^1(\mathcal{O}_Q(-n-1)) \to H^2(\mathcal{O}(-n-2)) \to \ldots .$$

The second sequence is easier to handle owing to the fact that $\mathcal{O}_Q$ refers to a subspace, whereas $\mathcal{O}^P$ refers to a factor space. There are three cases to consider, $Q_\alpha\,\bar{Q}^\alpha \gtreqless 0$. We have $H^p(\mathcal{O}_Q(m)) = H^p(Q^+,\mathcal{O}(m))$ where $Q^+ = Q \cap PT^+$, with $Q \subset PT$ given by $Q_\alpha Z^\alpha = 0$. When $Q_\alpha\,\bar{Q}^\alpha \geq 0$, Q^+ is Stein, so $H^1(\mathcal{O}_Q(-n-1)) = 0$ while $H^0(\mathcal{O}_Q(-n-1))$ is an infinite-dimensional space — given by the functional freedom of one holomorphic function of two variables, whence $\Phi_n \xrightarrow{Q_\alpha} \Phi_{n-1}$ is surjective:

$$Q_\alpha\,\bar{Q}^\alpha \geq 0\text{: } 0 \to C^{\infty 2} \to \Phi_n \xrightarrow{Q_\alpha} \Phi_{n-1} \to 0 \qquad \text{(exact)},$$

where $C^{\infty 2}$ symbolically stands for the freedom of one holomorphic function of two variables. When $Q_\alpha\,\bar{Q}^\alpha < 0$, then $Q^- = Q \cap PT^-$ is Stein whereas Q^+ has pseudoconcave boundary. Holomorphic functions on Q^+ extend to the whole of Q, so $H^0(Q^+,\mathcal{O}(m)) = H^0(Q,\mathcal{O}(m))$. This is non-vanishing only if $m \geq 0$ (polynomials of degree m) and then we have

$$0 \to H^0(PT,\ \mathcal{O}(m-1)) \to H^0(PT,\ \mathcal{O}(m)) \to H^0(Q,\ \mathcal{O}(m)) \to 0$$

$$0 \to H^0(PT^+,\mathcal{O}(m-1)) \to H^0(PT^+,\mathcal{O}(m)) \to H^0(Q^+,\mathcal{O}(m)) \underbrace{\to 0}_{\text{implied}} \qquad \text{(exact)}$$

whence (since $H^2(\mathcal{O}(-n-2)) = 0$ also), $\Phi_n \xrightarrow{Q_\alpha} \Phi_{n-1}$ is injective, with

$$Q_\alpha\,\bar{Q}^\alpha < 0\text{: } 0 \to \Phi_n \xrightarrow{Q_\alpha} \Phi_{n-1} \to C^{\infty 2} \to 0 \qquad \text{(exact)},$$

the functional freedom in $H^1(\mathcal{O}_Q(-n-1))$ being the same as in $H^0(\mathcal{O}_Q(-n-1))$ because of an argument involving Serre duality, and the relation between the cohomology of Q^+ and $\bar{Q}^-$.

The results for $\Phi_n \xrightarrow{P^\alpha} \Phi_{n-1}$ are similar and follow from the above by complex conjugation.

$$P^\alpha \bar{P}_\alpha > 0: \quad 0 \to \Phi_n \xrightarrow{P^\alpha} \Phi_{n+1} \to C^{\infty 2} \to 0 \qquad \text{(exact)},$$

$$P^\alpha \bar{P}_\alpha \leq 0: \quad 0 \to C^{\infty 2} \to \Phi_n \xrightarrow{P^\alpha} \Phi_{n+1} \to 0 \qquad \text{(exact)}.$$

Question: How does one characterize these maps *intrinsically*, so as to provide a "definition" of a twistor?

Acknowledgement: Thanks are due to Mike Eastwood.

REFERENCE

Penrose, R. 1975 *Twistor Theory, its Aims and Achievements.* In: *Quantum Gravity*, eds. C.J. Isham, R. Penrose & D.W. Sciama; pp. 267-407. Clarendon: Oxford.

104
SINGULARITIES AND TIME-ASYMMETRY

General relativity an Einstein centenary survey. Eds S. W. Hawking and W. Israel, pp. 581–638, Cambridge University Press, 1979

Various aspects of time symmetry and asymmetry, both classical and quantum mechanical, are considered at length and, I hope, with at least a little depth. In Section 12.2.7 I express some reasons for my dissatisfaction with Hawking's avowed belief that black holes and white holes are physically identical, and I give reasons for my own contention that space-time singularity structure must be constrained by something like a 'Weyl curvature hypothesis', implying that in some sense, quantum gravity must be a time-asymmetric scheme of things. On p. [630], there is a 'howler', where I have, in effect, forgotten the logarithm in Boltzmann's entropy formula (by dividing 10^9 by 10^{40} to get 10^{-31}, when I *should* have been dividing e^{10^9} by $e^{10^{40}}$ to get $e^{-10^{40}}$ (since the phase-space volumes are the exponentials of the entropy values)—the latter vastly increasing the strength of my case! (Let me also point out the irritating misprint 10 lines down on the final page [638], where 'exhibits' should be 'exhibit'.)

Singularities and time-asymmetry. *General relativity an Einstein centenary survey.*
Eds S. W. Hawking and W. Israel, pp. 581–638, Cambridge University Press, 1979

12. Singularities and time-asymmetry

R. PENROSE

12.1 Introduction

It has been a source of worry to many people that the general theory of relativity – that supremely beautiful description of the geometry of the world – should lead to a picture of spacetime in which singularities are apparently inevitable. Einstein himself had fought against the inevitability of such seeming blemishes to his theory, suggesting different possible ways out, of considerable ingenuity (e.g. the Einstein–Rosen bridge,[1] the attempt at a black-hole-avoiding stable relativistic star cluster,[2] the idea that a non-singular 'bounce' of the universe might be achieved through irregularities,[3] even his attempts at modifying general relativity to obtain a singularity-free unified field theory[4,5]). Yet, the researches of theorists in more recent years have driven us more and more in the direction of having to accept, and face up to, the existence of such singularities as true features of the geometry of the universe.

This is not to say that some mathematically precise concept of 'singularity' should now form part of our description of physical geometry – though much elegant work has been done in this direction in recent years. Rather, it seems to be that the very notion of spacetime geometry, and consequently the physical laws as we presently understand them, are limited in their scope. Indeed, these laws are even *self*-limiting, as the singularity theorems[6,7,8] seem to show. But I do not feel that this is a cause for pessimism. There is a need for new laws in any case; while, in my opinion, the presence and the apparent structure of spacetime singularities contain the key to the solution to one of the long-standing mysteries of physics: the origin of the *arrow of time*.

The point of view I am going to present does not stem fron any radical view of things. I shall adopt a basically conventional attitude on most issues – or so it would have seemed, were it not for the fact that my ultimate conclusions appear to differ in their basic essentials from those most commonly expressed on this subject! My arguments do not depend on detailed calculations, but on what seem to me to be certain 'obvious'

facts, whose very obviousness may contribute to their being frequently overlooked.

It was Einstein who was, after all, the supreme master at deriving profound physical insights from 'obvious' facts. I hope I may be forgiven for trying to emulate him on his hundredth birthday, for I feel sure that he would have cared not one fig for the supposed significance of so arbitrary an anniversary!

12.2 Statement of the problem

The basic issue is a familiar one.[9,10,11] The local physical laws we know and understand are all symmetrical in time. Yet on a macroscopic level, time-asymmetry is manifest. In fact, a number of apparently different such macroscopic arrows of time may be perceived. To these may be added the only observed time-asymmetry of particle physics – which features in the decay of the K^0-meson. And there is one further related issue, namely the interpretation of quantum mechanics, which I feel should not be banished prematurely from our minds in this connection. The conventional wisdom has it that, despite an initial appearance to the contrary, the framework of quantum mechanics contains no arrow.[9,12–16] I do not dispute this wisdom, but nevertheless believe, for reasons that I shall indicate, that the question must be kept alive.

Let me list, therefore, seven apparently independent arrows (or possible arrows) that have been discussed in the literature:[9] section 12.2.1, K^0-meson; section 12.2.2, quantum-mechanical observations; section 12.2.3, general entropy increase; section 12.2.4, retardation of radiation; section 12.2.5, psychological time; section 12.2.6, expansion of the universe; and section 12.2.7, black holes versus white holes. I shall discuss each of these in turn.

12.2.1 Decay of the K^0-meson

Can the asymmetry that is present in the decay rate of the K^0-meson have any remote connection with the other arrows? The effect, after all, is utterly minute. The T-violating component in the decay is perhaps only about one part in 10^9 of the T-conserving component[17–21] – and, in any case, the presence of this T-violation has to be inferred, rather than directly measured, from the presence of a minute CP-violation ($\sim 10^{-9}$) together with the observation that CPT-violation, if it exists, must be even smaller in this interaction ($\ll 10^{-9}$). A very weak interaction indeed

Singularities and time-asymmetry. *General relativity an Einstein centenary survey.*
Eds S. W. Hawking and W. Israel, pp. 581–638, Cambridge University Press, 1979

(or very weak component of a weak interaction) seems to be involved, and it plays no significant role in any of the important processes that govern the behaviour of matter as we know it. Gravitation, of course, is even weaker, and does, in fact, dominate the motion of matter on a large scale. But the differential equations of general relativity are completely time-symmetric as are Maxwell's equations and, apparently, the laws of strong interactions and ordinary weak interactions.

Yet the tiny effect of an almost completely hidden time-asymmetry seems genuinely to be present in the K^0-decay. It is hard to believe that Nature is not, so to speak, 'trying to tell us something' through the results of this delicate and beautiful experiment, which has been confirmed several times.[20] One of the suggestions that was put forward early on was that the T-violating effect arose via some cosmological long-range interaction, whereby matter-tô-antimatter imbalance provided the required asymmetry.[50] But subsequent analysis[19,20] has rendered this viewpoint implausible. It seems that the asymmetry is really present in the *local* dynamical laws. This is a matter that I shall return to later. I believe that it is a feature of key significance.

12.2.2 Quantum-mechanical observations

In standard quantum mechanics, the dynamical evolution of a state takes place according to Schrödinger's equation. Under time-reversal this equation is transformed to itself provided i is replaced by $-$i. But Schrödinger's equation must be supplemented by the procedure ('collapse of the wavefunction') whereby the current state vector is discarded whenever an 'observation' is made on a system, and is replaced by an eigenstate ψ_Q of the Hermitian operator Q which represents the observable being measured. As it stands, this procedure is time-asymmetric since the state of the system is an eigenstate of Q just after the observation is made, but not (normally) just before (figure 12.1(a)). However, this asymmetry of description is easily remedied:[14] in the time-reversed description, one simply regards this same eigenstate ψ_Q as referring, instead, to the state just before the observation and Schrödinger's equation is used to propagate backwards until the previous observation (with operator P) is reached. Thereupon this (backwards-evolved) state vector is discarded (according to a time-reversed version of the 'collapse of the wavefunction') and an eigenstate ψ_P of P (corresponding to the actual result of the observation P) is substituted (figure 12.1(b)).

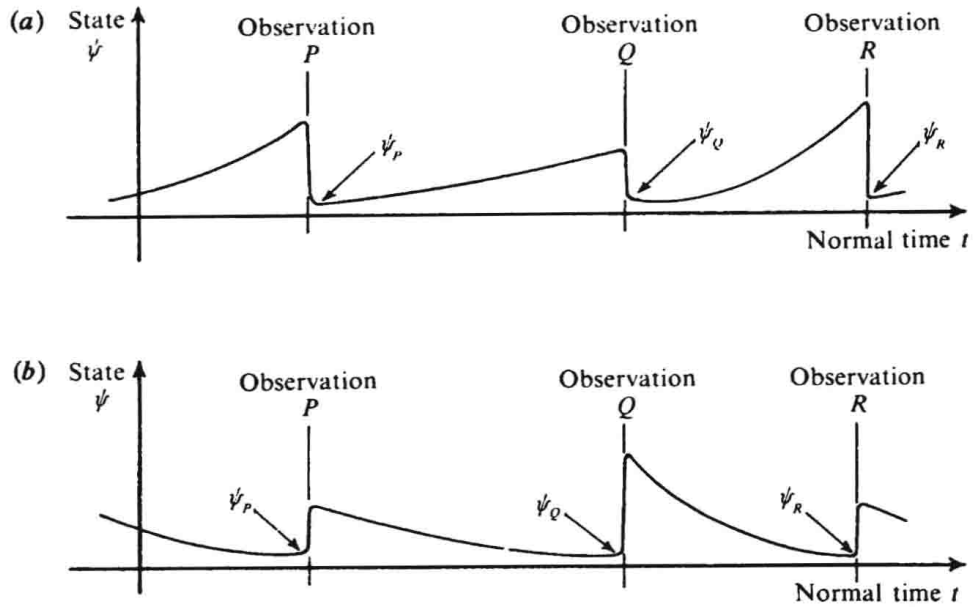


Figure 12.1. (*a*) Conventional (Schrödinger) picture of development of a wavefunction. (*b*) Essentially equivalent 'time-reverse' development of a wavefunction.

The relative probability of the observed Q-value, given the observed P-value, or of the observed P-value, given the observed Q-value, is the same in each mode of description, being

$$|\langle\psi_Q|U\psi_P\rangle|^2 = |\langle\psi_P|U^{-1}\psi_Q\rangle|^2,$$

where U is the unitary operator representing the evolution (according to Schrödinger's equation) of a state from the time of P to the time of Q. Since these probabilities are the only observational manifestation of the wavefunction ψ in any case, we see that the two modes of description are equivalent, and that the framework of quantum mechanics is time-symmetric.

It is often stated[16,22] that the actual value of the wavefunction at any time is not properly a description of physical reality. This is strikingly illustrated by figure 12.1, if time-symmetry is to be a feature of microscopic physics. But in any case there is a well-known difficulty, even with special relativity, concerning taking too strong a view that the wavefunction describes physical reality. For an 'observation' would seem to collapse the wavefunction into one of its eigenstates simultaneously with that observation – where 'simultaneous' presumably refers to the rest-frame of the one making the observation. This can lead to conceptual problems, when two spacelike-separated observations are carried out,

concerning the question of the *ordering* in which two collapses of the wavefunction take place. The difficulty is pinpointed particularly well in the famous thought experiment of Einstein, Podolsky and Rosen.[22]

There are, however, *other* situations in which it seems hard to maintain the view that the wavefunction (or state vector) does *not* give a proper description of physical reality in accordance with figure 12.1(*a*). Consider an isolated system on which observation P has just been made, giving a *conventional* description of a state ψ_P (eigenstate of P) which evolves forwards in time according to Schrödinger's equation to give a state $U\psi_P$ at some later time. Now (according to the conventional framework of quantum-mechanics, and assuming that no additional principles are incorporated) the operator UPU^{-1} is just as 'good' an observable as P. Furthermore, provided that the eigenvalue λ corresponding to ψ_P is simple, $U\psi_P$ has the property that it (or a nonzero multiple of it) is the *unique* state for which the probability is unity of giving the value λ for the observation UPU^{-1}. The isolated system cannot (in the ordinary way of looking at things) 'know' that the observation UPU^{-1} may be about to be performed upon it, but it must be prepared for that eventuality! So it seems that the information of $U\psi_P$ (up to phase) must be stored in the system, i.e. that the wavefunction *does* describe physical reality.

Of course UPU^{-1} may correspond to an utterly outlandish and totally impracticable experiment, as, for example, if in the case of the Schrödinger 'cat paradox'[24] we were to attempt to verify a resulting state $U\psi_P$: 'complex linear combination of dead cat and alive cat'. The very outlandishness of such an experiment suggests that $U\psi_P$ (and UPU^{-1}) may not, after all, 'really' refer to reality! But that is my whole point. There is something missing or something inappropriate about the laws of quantum mechanics, when applied to such situations. There is also something absurd about the whole idea of a collapsing wavefunction (or of any of the other essentially equivalent alternative ways of describing the same phenomenon, such as the conscious observer threading his way through an Everett-type[26–29] many-sheeted universe) as a description of physical reality. Yet what is a physical theory for if not to describe reality? In this I feel that I must align myself with many of the originators and main developers† of quantum mechanics – and, not least, with Einstein himself[23] – in believing that the resolution of the question of 'observations' is not to be found within the formalism of quantum mechanics itself. Some new (presumably nonlinear) theory seems to be required in

† E.g., in their different ways, Bohr,[30] Schrödinger,[24] Dirac,[31] Wigner.[25]

which quantum mechanics and classical mechanics each emerges as a separate limiting case.

The issue has importance here for two reasons. In the first place, the making of an observation seems to be associated with an *irreversible* process, depending upon an essential entropy increase. It is not obvious that what is missing (or wrong) about quantum mechanics is not some fundamentally *time-asymmetric* law. So the demonstration of time-symmetry in the formalism of quantum mechanics does not really settle the question of time-(a)symmetry in quantum-mechanical observations.

The second reason why the issue is important here has to do with the question of the role of quantum gravity. One must bear in mind the possibility that it might be the presence of a (measurable) *gravitational field* that takes the description of a physical system out of the realm of pure quantum physics.[32,33] And if a *new law* is needed, especially to cover situations in which both quantum and classical physics are being stretched in the extreme, then quantizers of gravity, beware!

I shall return to these questions in section 12.4. But, for the present, I propose to be wholly conventional in my attitude to quantum mechanics: it contains no manifest arrow, and the solution to the problem of macroscopic time-asymmetry must be sought elsewhere.

12.2.3 General entropy increase

The statistical notion of entropy is, of course, crucial for the discussion of time-asymmetry. And if the (important) local laws are all time-symmetric, then the place to look for the origin of statistical asymmetries is in the boundary conditions. This assumes that the local laws are of the form that, like Newtonian theory, standard Maxwell–Lorentz theory, Hamiltonian theory, Schrödinger's equation, etc., they determine the evolution of the system once boundary conditions are specified, it being sufficient to give such boundary conditions *either* in the past *or* in the future. (The boundary values are normally specified on a spacelike hypersurface.) Then the statistical arrow of time can arise via the fact that, for some reason, the *initial* boundary conditions have an overwhelmingly *lower* entropy than do the *final* boundary conditions.

There are several issues that must be raised here before we proceed further. In the first place, there is something rather unreasonable about determining the behaviour of a system by specifying boundary conditions at all, whether in the past *or* future. The 'unreasonableness' is particularly apparent in the case of future boundary conditions. Suppose I throw my

watch on a stone floor so that it shatters irreparably, and then wait for 10 minutes. The future boundary condition is a mess of cogs and springs, but with minutely organized velocities of such incredible accuracy that when reversed in direction (i.e. with the clock run backwards) they suddenly reassemble my watch after a 10-minute period of apparent motionlessness. Though the models of physics that we are using (e.g. Newtonian theory) allow, in principle, such accuracy to be defined, we do not know (nor do we demand) that our models of physical laws correspond with such precision to reality. The problem is there also even for past boundary conditions, as stressed by Born.[34] And Feynman[35] has pointed out that, in Newtonian theory, if all the positions and velocities of a complex system are known to a certain accuracy, then all the accuracy is lost in less time than it takes to state that accuracy in words! Born (and Feynman) invoke this argument to demonstrate that classical mechanics is, in a sense, no more deterministic than quantum mechanics. Of course quantum mechanics has the additional problem of what happens when an 'observation' is made – and 'observations' seem to be necessary, in the normal view, to keep the wavefunctions from spreading throughout space.

I mention such things mainly to point out difficulties. But I shall ignore them henceforth and follow the conventional path that boundary conditions work! There is, however, a somewhat related question that needs further comment. Consider, again, my shattered watch as a future boundary condition. It will be seen that though this state has a higher entropy than that before my watch was shattered – and, therefore, in the normal way of looking at things, is a less 'unusual' state than the earlier one – the later state is nevertheless a very strange one indeed when one comes to examine it in minute detail, in view of the very precise correlations between the particle motions. But again I shall adopt a conventional 'macroscopic' view here. This strangeness is not of the kind that is described as 'low entropy'. And the 'reason' that I had a watch earlier is not that these precise correlations exist in the *future* boundary conditions, but that something in the *past* (say a watch factory) had a lower entropy than it might otherwise have had. Likewise, the 'reason' for the precise correlations in the particle motions of the shattered watch can be traced back to the factory, not the other way around.

I do not feel that this is begging the question of time-asymmetry. It could perfectly well have been the case, in a suitably designed universe, that some processes behave like my watch, while others (using the time-sense defined by my watch while it was still working!) indulge in apparently miraculous assembly procedures which suggest that special

(low-entropy) *future* boundary conditions should be invoked to provide the 'reason' for *their* behaviour. But our universe seems not to work that way.

An important related question that I have glossed over so far is that of coarse-graining.[36] What does entropy mean anyway? Is it a definite physical attribute of a system which, like energy–momentum, seems to be independent of the way that we look at it? In practice, entropy can normally be treated that way (for example, in physical chemistry). But for the most general definitions of entropy that might be expected to apply to a complicated system such as a watch, we need some apparently rather arbitrary (i.e. non-objective) way of collecting together physical states into larger classes (coarse-graining) where the members of each class are considered to be indistinguishable from one another. The entropy concept then refers to the classes of states and not to the individual states. Then, the (Boltzmann) entropy of a class containing N distinct (quantum) states is

$$S = k \log N$$

(where k is Boltzmann's constant). In fact, several conceptually different definitions of entropy are available.[36–38] But the whole question is clearly fraught with difficulties.† (The phenomenon of 'spin-echo' is one striking example that emphasizes these difficulties.[39] I am not even convinced that 'entropy increase' is at all an appropriate concept for describing the shattering of my watch. Probably taking a bath increases the entropy enormously more – while, in the case of my watch, the proportional increase in entropy must be quite insignificant.) The question of the objectivity of entropy will be returned to in section 12.4. But for the moment I hastily take refuge once more in conventionality: entropy is a concept that may be bandied about in a totally cavalier fashion!

12.2.4 Retardation of radiation

The question of boundary conditions is also intimately involved in the next of our arrows, that of retarded radiation. We may separate this phenemenon into two quite distinct aspects: the entropy question again, and the question of source-free or sink-free radiation. Retardation is not just a feature of electromagnetic radiation, of course, though it is usually

† I am leaving aside also such important questions as the 'H-theorem'[36–38] and 'branch systems'.[10,11] They do not *explain* the time-asymmetric origin of the total entropy imbalance.

discussed in that context. Imagine a stone thrown into a pond. We expect to see ripples expanding outwards from the point of entry to have their energy gradually dissipated, especially when they hit the bank. We do not expect to see, before the stone reaches the water, ripples being produced at the bank with such precise organization that they converge upon the point of entry at the exact moment that the stone enters the water. Still less do we expect to see such ripples converging inwards from the bank to some point in the middle of the pond, at which they entail the sudden ejection of a stone into the air! Such extraordinary behaviour is perfectly in agreement with the local physical laws. But its occurrence would require the sort of precise correlations in particle motions that could only be explained by some low-entropy future boundary condition.

I should emphasize, once more, a point made in the last section, since I feel that it is a key issue: correlations are present in the detailed particle motions in the future *because* the entropy was low in the past. Likewise, similar (but time-reversed) correlations are absent in the past *because* the entropy is high in the future. The latter statement is an unusual form of words, but I am trying to be unbiased with respect to time-ordering.† My point of view is that the correlations are not to be viewed as the 'reason' for anything; but low entropy (itself to be explained from some other cause) *can* provide the 'reason' for the correlations. (This way around we avoid the problem of over-extreme precision in physical laws.) And once more I stress that the 'specialness' of a state, due to its possessing intricate particle correlations of this kind, is *not* the type of 'specialness' that *at that time* corresponds to a low entropy. That is an essential point of coarse-graining.

So we see that the normal retarded behaviour of the ripples correspond to low entropy in the past and correlations in the future, while the two situations I have described, which seem to involve *advanced* behaviour of the ripples, involve some very precise correlations in the past of the kind leading to a reduction in entropy. Furthermore, an alternative hypothetical *retarded* situation, in which a stone is suddenly ejected from the pond accompanied by ripples propagating outwards towards the bank,‡ also involves such precise correlations (this time in the motions of the

† This leads to a logical reversal of the viewpoint expressed by O. Penrose and Percival[40] in their 'law of conditional independence', according to which the absence of initial correlations is *postulated*. The world-view of section 12.3.3 provides a certain justification for this law.

‡ The reader may notice the close relation between this situation and the 'zag' motion described by Gold.[9,41] Likewise, the converging waves meeting the falling stone are 'zaglike', while the other two are 'ziglike'.

particles near the stone at the bottom of the pond). Thus, in these situations we have no need to invoke any extra hypothesis to explain why the ripples are retarded. The entropy hypothesis is already sufficient to rule out the two advanced situations as overwhelmingly improbable – and it also rules out the above unreasonable retarded situation with the ejected stone (assuming the absence of any other agency responsible for ejecting the stone, such as an underwater swimmer, etc.).

The situation with electromagnetic radiation is, for the most part, similar to that of the ripples. There is a minor difference here, however, in the case of stars shining in a largely empty universe: it might well be that some of this radiation is never absorbed by any matter, but continues on indefinitely as the universe expands, or else terminates its existence on a spacetime singularity; likewise, there might be source-free radiation present in the universe, which had been produced directly in the big bang (or in a white hole) or possibly had come in from infinity from a previously collapsing phase of the universe.

I do not believe that these possibilities really make any essential difference to the discussion. I mention them mainly because a great deal has been written on the subject of 'the absorber theory of radiation'. In this view,[42,43] the contribution to the electromagnetic field due to each charged particle is taken to be half advanced and half retarded, while any additional source-free or sink-free radiation is regarded as 'undesirable'. By postulating the absence of such additional radiation, a link between the expansion of the universe and the retardation of radiation is obtained – though, in my opinion, not very convincingly. (And I have to confess to being rather out of sympathy with the whole programme, which strikes me as being unfairly biased against the poor photon, not allowing it the degrees of freedom admitted to all massive particles!)

In any case, the relevance of the entropy argument to the question of retardation seems to me to be quite independent of this.[44] The presence of free radiation coming in from infinity (or from the big bang singularity, say) which converges on a searchlight the moment it is switched on – or some other such absurdity – corresponds just as much to entropy-decreasing-type correlations in the initial state as would radiation coming in from sources. The only difference is that the correlations are just put directly into the photons themselves rather than into the particles producing the photons. So such correlations would be expected to be absent if the entropy in the future is to be high. Correspondingly, there is no objection to such correlations (in time-reversed form) being present in the *future* boundary conditions, because the entropy was low in the

past – and this is, of course, necessary in order that the stars should shine.

The reader may be concerned about how one actually specifies boundary conditions at infinity, or on a spacetime singularity, in order to discuss such correlations in any detail. Of course serious technical problems can arise, particularly in the case of singularities. But the details of these problems should not substantially affect the foregoing discussion – at least if cosmic censorship holds true. I prefer to postpone these questions until section 12.3.2 except just to mention that under certain circumstances (e.g. in a big bang model in which the total charge within some observer's past light-cone is nonzero) there is necessarily a certain amount of source-free radiation present (and in other circumstances, a certain amount of sink-free radiation).[45] There is no reason to believe that this radiation should be correlated in any way which is incompatible with the entropy arguments. The stars still shine 'outwards', rather than 'inwards', whether or not there is some additional radiation permeating space – provided that this radiation has less than the intensity of a star and that it is not specially correlated.

12.2.5 Psychological time

The arrow most difficult to comprehend is, ironically, that which is most immediate to our experiences, namely the feeling of relentless forward temporal progression, according to which potentialities seem to be transformed into actualities. But since the advent of special relativity it has become clear that at least in *some* respects this feeling is illusory. One has the instinctive (or perhaps learnt) impression that one's own concept of temporal progression is universal, so that the transformation of potentialities into actualities that each one of us feels to be taking place ought to occur simultaneously for all of us. Special relativity tells us that this view of the world is false (and it is this lesson that had probably represented the major obstacle to the understanding and acceptance of the theory). Two people amble past one another in the street. What potential events are then becoming actual events on some planet in the Andromeda galaxy? According to the two people, there could be a discrepancy of several days!† (And adopting a view that events are becoming actual on, say, one's past or future light-cone – rather than

† To make the question seem more relevant, imagine that 'at that very moment' a committee is sitting on the planet, deciding the future of humanity!

using Einstein's definition of simultaneity – only makes the subjectivity of these occurrences even worse!)

So relativity *seems* to lead to a picture in which 'potentialities becoming actualities' is either highly subjective or meaningless. Nevertheless the feeling remains very strong within us that there is a very fundamental difference between the past and the future, namely that the past is 'actual' and unchangeable, whereas the future can yet be influenced and is somehow not really fixed. The usual view of the world according to relativity denies this, of course, presenting a rigid four-dimensional determinate picture and telling us that our instinctive feelings concerning the changeability of the future are illusory.

But I do not think that we should just dismiss such feelings out of hand. It is possible to envisage model universes in which the future is yet indeterminate, while the past is fixed. Imagine a continually branching universe, like that depicted in figure 12.2. One is to depict oneself located

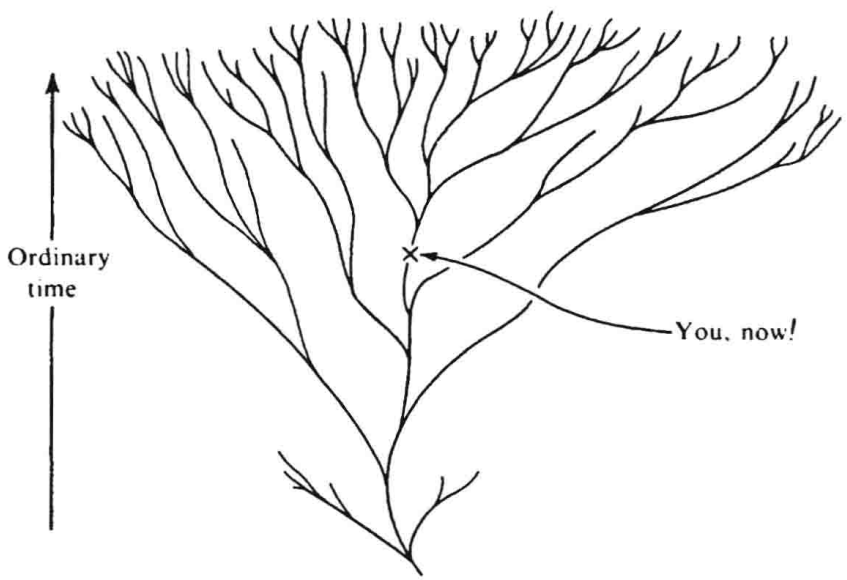


Figure 12.2. A model universe, branching into the future.

at some universe-point in the midst of it all, 'moving' up the picture as one's psychological time unfolds. In this model, the branching takes place only into the future. The path into the past from that point (i.e. the past history of the universe) is absolutely unique, whereas there are many alternative branches into the future (i.e. many alternative possible future histories for the universe, given the present state).

There are, in fact, (at least) two ways to make such a model relativistic. In the first (figure 12.3(*a*)), the branching takes place along the future

 Singularities and time-asymmetry. *General relativity an Einstein centenary survey.* Eds S. W. Hawking and W. Israel, pp. 581–638, Cambridge University Press, 1979

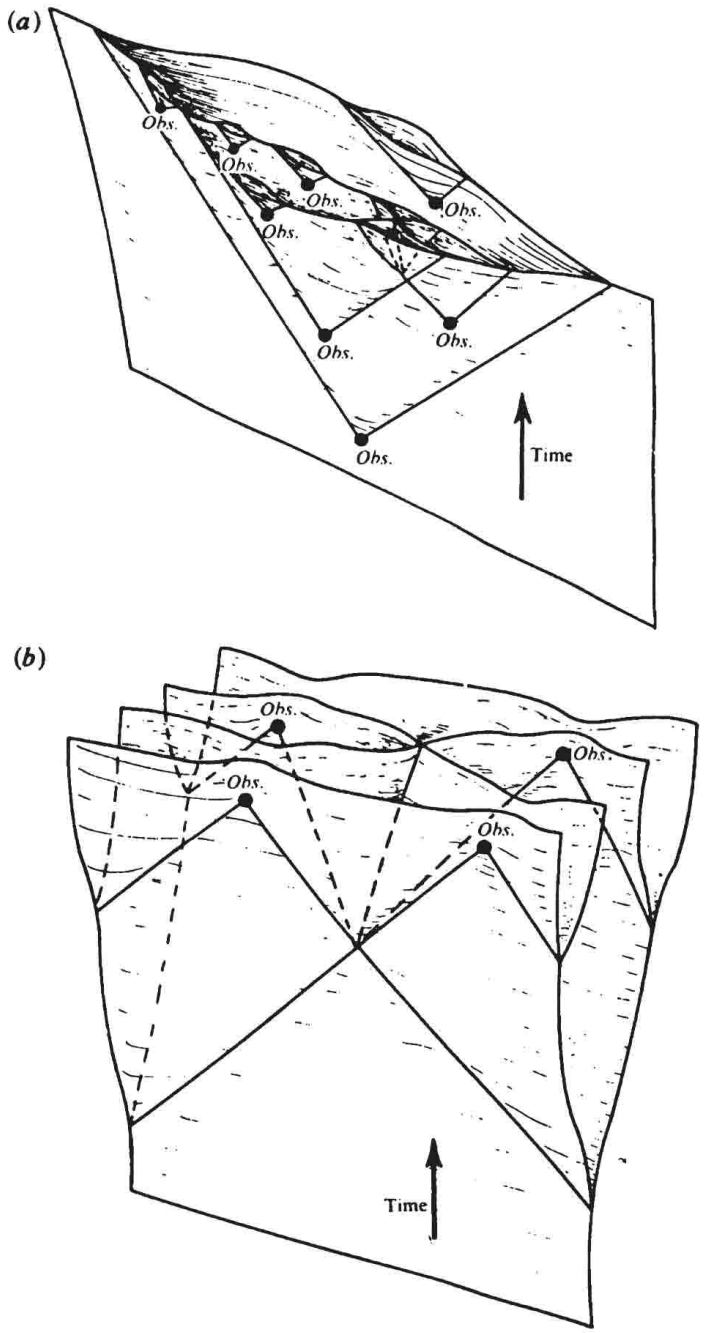


Figure 12.3. Two possible versions of a relativistic future-branching universe: *Obs.* means event at which an 'observation' is made.

light-cones of the points at which 'observations' (presumably quantum-mechanical) are made, and in the second (figure 12.3(*b*)), along the past light-cones of such points. This second case would seem to have considerably less plausibility than the first, since the universe has to 'know' that it has to branch in advance of the observation – which suggests

the presence of some sort of precise correlations in the past, like those leading to advanced-type radiation. The first model, however, is not altogether implausible. It is possible to envisage, for example, that the branching accompanies a kind of retarded collapse of the wavefunction, where on each branch the wavefunction starts out as a different eigenvector of the operator representing the observation.

Such a model may be referred to as an 'Everett-type' universe,[26–29] although it is by no means clear to me that this is really the kind of picture that the Everett formalism of quantum mechanics leads to. For, in the first place, if the Everett picture is to be essentially a reformulation of standard quantum mechanics, it ought to be time-reversible, i.e. there ought to be as much branching into the past as into the future. In the second place, in the Everett picture, one envisages a single wavefunction for the whole universe, which never itself 'collapses'. Instead, it becomes naturally represented as a linear combination of states, in each of which is a measuring apparatus and a physical system appearing to be in a separate eigenstate of the apparatus. Possibly, and with further assumptions as to, say, the nature of the Hamiltonian involved, etc., an Everett picture could, to a certain degree of approximation, be shown to lead to a picture somewhat resembling figures 12.2 and 12.3. But no-one appears to have done this. In particular, without some grossly time-asymmetrical additional assumption there would be nothing to rule out the time-reverses of figures 12.2 and 12.3, or innumerable possibilities in which branching occurs both to the past and to the future, with many branches temporarily separating and then coming together again.

Such considerations are clearly wildly speculative. But one is, in any case, groping at matters that are barely understood at all from the point of view of physics – particularly where the question of human (or non-human) 'consciousness' is implicitly involved, this being bound up with the whole question of psychological time, though not necessarily with the Everett picture. But spacetime models resembling figure 12.3(*a*), say, might well be worthy of study. They could be legitimate mathematical objects, e.g. four-dimensional Lorentzian manifolds subject to Einstein's field equations (say), but where the Hausdorff condition is dropped.[46] It could be argued that such a model is more in accordance with one's intuitive feelings of a determinate past and an indeterminate future than is our normal picture of a Hausdorff space-time. And one could claim that our feeling that 'time moves forwards into the future', rather than 'into the past', is natural in view of the fact that in the 'forward' direction potentialities become actualities, rather than the other way about! Such a

model could be viewed as an 'objective' description of a world containing some strongly 'subjective' elements. One could envisage different conscious observers threading different routes through the myriads of branches (either by chance, say, or perhaps even by the exercise of some 'free will'). And each such observer would have a different 'subjective' view of the world.

This is not to say that I have any strong inclinations to believe in such a picture. I feel particularly uncomfortable about my friends having all (presumably) disappeared down different branches of the universe, leaving me with nothing but unconscious zombies to talk to! I have, in any case, strayed far too long from my avowed conventionality in this discussion, and no new insights as to the *origin* of time-asymmetry have, in any case, been obtained. I must therefore return firmly to sanity by repeating to myself three times: 'spacetime is a *Hausdorff* differentiable manifold; spacetime is a *Hausdorff*...'!

I have as yet made no real attempt to relate the *direction* of psychological time to the key question of entropy. It is clear that any attempted answer must necessarily be very incomplete, in view of our very rudimentary understanding of what constitutes psychological time. And I should emphasize that it is not just a question of the past being (apparently) more certainly *knowable* than the future. Indeed, it is not always the case that the past is easier to ascertain than the future, i.e. that 'retrodiction' is more certain than 'prediction'. At a meeting in Cornell in 1963, Gold[41] pointed out that it was far easier to predict the future behaviour of a recently launched Soviet satellite than to ascertain the time and place of launching! Furthermore, it is not very clear why the phenomenon of increasing entropy should, in any case, be compatible with ease in retrodiction and difficulty in prediction. If earlier states are more 'exceptional' and later ones more 'usual', then it would seem that we should be on safer ground predicting later ones than retrodicting earlier ones! Since, in practice, retrodiction is normally easier (or, at least, more accurate) than prediction (memory being more reliable than soothsaying), the precise relation between this and the entropy question is obscure.

But the issue is really rather different from this. It is not the ease in inferring the past that is relevant here, but the feeling that the past is unchangeable. Likewise, it is not the difficulty that we might have in guessing (or trying to calculate) the future that concerns us, but the feeling that we can affect the future. Thus, despite Gold's observation that the Soviet satellite's future was accurately predictable, one might have the

lingering unease that (had the technical expertise been available at that time) someone might have tried to intercept it. On the other hand, it seems totally inconceivable that any action could be applied *after* the satellite was in orbit to change the launching date! But if the future is to be, in principle, not essentially different from the past, one must entertain the awesome possibility[47] that even the 'fixed' past might conceivably, under suitable circumstances, be changeable! Is it just a question of 'money' that prevents this? (The science-fiction possibility suggests itself of a powerful government going one stage further than falsifying history: namely, actually *changing* the past!) I prefer to leave this question well alone.

There is a rather more helpful way of looking at the intuitive past–future distinction, however. One tends to view events in the past as providing 'causes' or 'reasons' for events in the future, not the other way around. This, at least, *is* compatible with the views I have been presenting in sections 12.2.3 and 12.2.4. My attitude has been that a low entropy at one time may be regarded as providing a 'reason' for precise correlations in particle motions at another, but that the presence or absence of such correlations should not be regarded as providing 'reasons' for anything else. We observe low entropy in the past and infer precise correlations in the future. Thus it is the presence of low entropy in the past that implies that the state of the past provides 'reasons' for the state of the future, not the other way around. This seems to me to be wholly sensible. If it had been the case in our world that collections of broken cogs and springs would sometimes spontaneously assemble themselves into working watches, then people would surely not be averse to attributing the 'causes' of such occurrences to events in the future. Such occurrences might co-exist with others of the more familiar kind, whose 'causes' could be attributed to events in the past. But our universe happens not to be quite like that! The 'causes' of things in *both* types of universe would be traced to situations of low entropy. And in *our* universe these low-entropy states turn out to be in the past.

So at least in this case our psychological feeling of a distinction between past and future can be directly linked to the entropy question. Perhaps the other aspects can too.

12.2.6 Expansion of the universe

I have implicitly indicated in section 12.2.4 that the expansion of the universe should not be regarded as directly responsible for the retar-

dation of radiation, the latter phenomenon being simply one of the many consequences of an assumption that the initial state of the universe was of a far lower entropy than its final state will be (and, correspondingly, that entropy-decreasing correlations were absent in the initial state). I now wish to argue that the expansion of the universe cannot, in itself, be responsible for this entropy inbalance either.

For let us suppose that the contrary is the case and that, for some reason, increasing entropy is a necessary concomitant of an expanding universe. By time-reversal symmetry, this view would entail, correspondingly, that in a contracting universe the entropy should decrease.[48] There are two main situations to consider. First, it might be that the expansion of our actual universe will some day reverse and become a contraction, in which case, according to this view, the entropy would start decreasing again to attain a final low value. The second possibility is that the expansion will continue indefinitely and that the entropy will likewise continue increasing for ever – until a maximum entropy state is reached (ignoring the question of Poincaré cycles, etc.).

It seems that there are very serious objections to the idea that the trend of increasing entropy will reverse itself when the universe reaches maximum expansion. It is hard to see how such reversal could take place without some sort of thermal equilibrium state having been reached in the middle. Otherwise one would have to envisage, it seems to me, a middle state in which phenomena of the normal sort (e.g. retarded radiation and shattering watches) would co-exist with phenomena of the 'time-reversed' sort (e.g. advanced radiation and self-assembling watches). While it is possible to contemplate such situations 'for the purposes of argument', it is a different matter altogether for us to take them seriously for our *actual* universe. Furthermore, the moment of time-symmetry would be reached, it would seem, whilst normal retarded light from very distant galaxies is still coming in (those distant galaxies appearing still to be receding) and, at the same time, there would be advanced light behaving in the time-reversed way (i.e. specially correlated and converging on approaching galaxies). There would seem to be some serious self-consistency problems here[9] (though I am not claiming that they are totally insurmountable). I cannot find it in myself to take such a picture seriously – though some others have apparently not found their intuitions to be so constrained![48]

We might suppose, on the other hand, that the timescale for the reversal of the expansion is so enormously long that an effective thermal equilibrium can be achieved at maximum expansion. But such times that

one must contemplate for this are of a completely different order from the normal cosmological scales. *In effect*, then, such a universe does not recontract at all, and the situation can be considered alongside that of the indefinitely expanding universe-models.

One might think that these models would avoid the problems just considered, but this is not so. Let us envisage an astronaut in such a universe who falls into a black hole. For definiteness, suppose that it is a hole of $10^{10} M_\odot$ so that our astronaut will have something like a day inside, for most of which time he will encounter no appreciable tidal forces and during which he could conduct experiments in a leisurely way. In figure 12.4 the situation is depicted in a standard conformal diagram

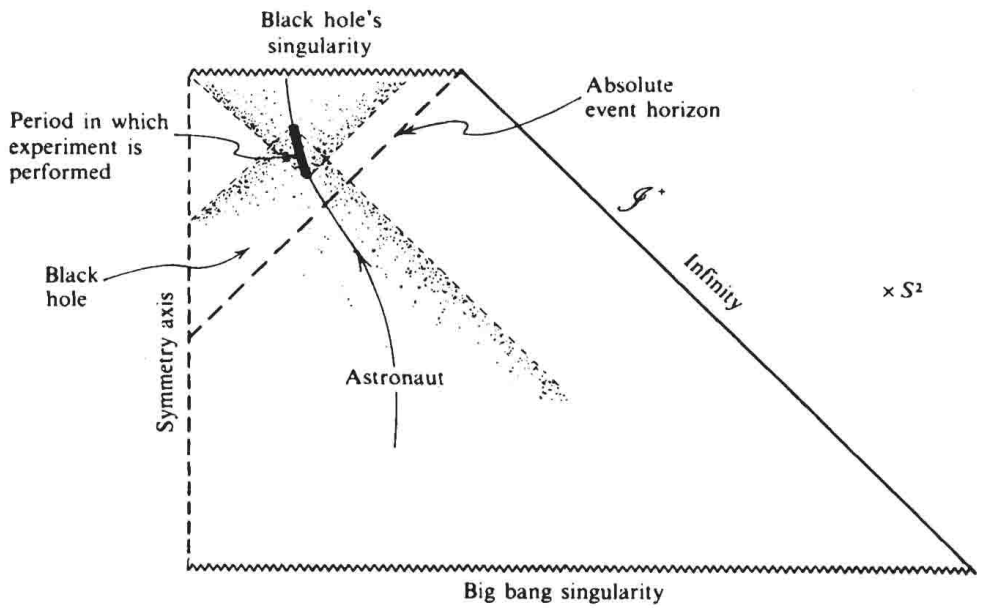


Figure 12.4. Conformal diagram of astronaut falling into black hole in Friedmann $k = -1$ universe.

(light-cones drawn at 45° and spherical symmetry assumed). Notice that the entire history of the astronaut beyond his point of crossing the absolute event horizon lies within the past domain of dependence of the black hole's singularity – and also within the future domain of dependence of the big bang singularity. Suppose that experiments are performed by the astronaut for a period while he is inside the hole. The behaviour of his apparatus (indeed, of the metabolic processes within his own body) is entirely determined by the conditions at the black hole's singularity (assuming that behaviour is governed by the usual hyperbolic-type differential equations) – as, equally, it is entirely determined by the

conditions at the big bang. The situation inside the black hole differs in no essential respect from that at the late stages of a recollapsing universe. If one's viewpoint is to link the local direction of time's arrow directly to the expansion of the universe, then one must surely be driven to expect that our astronaut's experiments will behave in an entropy-*decreasing* way (with respect to 'normal' time). Indeed, one should presumably be driven to expect that the astronaut would believe himself to be coming *out* of the hole rather than falling in (assuming his metabolic processes could operate consistently through such a drastic reversal of the normal progression of entropy).

I have to say that I cannot help regarding this possibility as even *more* of an absurdity than the entropy reversal at maximum expansion of a recollapsing universe! One would presumably not expect the entropy to reverse suddenly as the astronaut crosses the horizon (and thereafter, the collapsing aspect of his boundary conditions *enormously* outweighs the expanding one). So, strange behaviour in the entropy would have to occur well before the astronaut actually crosses the horizon – whereupon the astronaut could change his mind and accelerate outwards, so avoiding capture by the hole and being in a position to report his strange findings to the outside world! Indeed, the black hole argument can also be applied in the recollapsing universe. We do not need to wait for the whole universe to recollapse in order for the absurdities of this viewpoint to manifest themselves. (It may be that some holders of this viewpoint have a disinclination to accept the reality of black holes in any case. I have no desire to enter into the arguments – in my view very compelling – in favour of black holes here, but refer the reader to the literature.[7.51])

An argument could be put forward that the spacetime depicted in figure 12.4 is based on a too strong and unrealistic assumption of spherical symmetry. In fact this is not really the case; the dropping of spherical symmetry should make no essential qualitative difference to the picture, provided only that a (suitably strong) assumption of cosmic censorship is made. I prefer to postpone a more detailed discussion of this point until section 12.3.2. For the moment it is sufficient that I may fall back on conventionality again to conclude that the expansion of the universe is *not, in itself*, somehow responsible for the fact that the entropy of our universe is increasing.

This is not to say that I regard the correspondence between these two awesome facts as entirely fortuitous. Far from it. For I shall argue later on that both are consequences of the very special nature of the big bang – a special nature that is *not* to be expected in the singularities of recollapse.

12.2.7 Black holes versus white holes

General relativity is a time-symmetric theory. So, to any solution of its equations (with time-symmetric equations of state) that is asymmetric in time, there must correspond another solution for which the time-ordering is reversed.† One of the most familiar solutions is that representing (spherically symmetric) collapse of a star (described using, say, the T_{ab} of 'dust') to become a black hole.[54,55,7] In time-reversed form this represents what is referred to as a 'white hole' finally exploding into a cloud of matter. Spacetime diagrams for the two situations are given in figure 12.5.

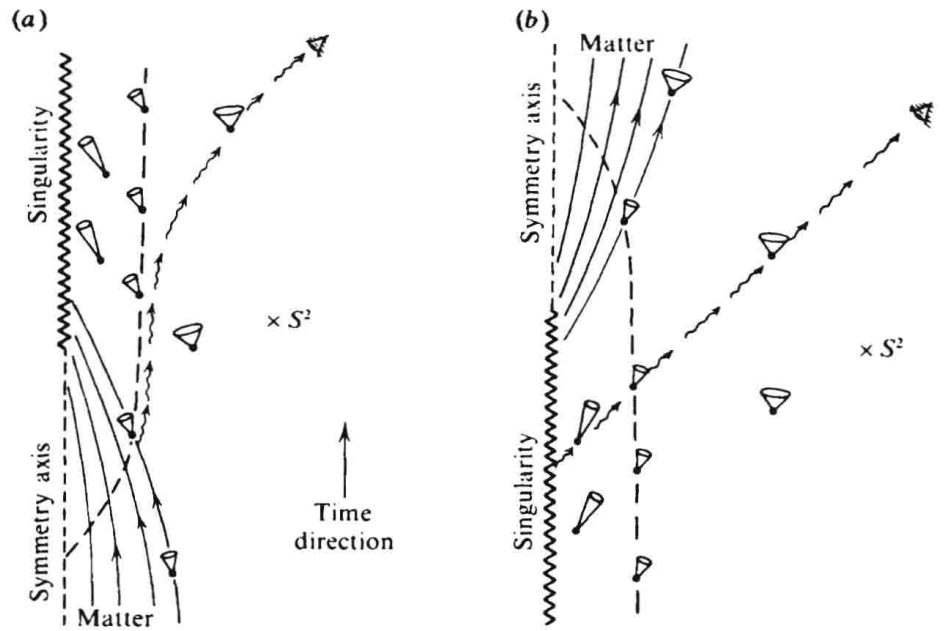


Figure 12.5. Black and white holes (Finkelstein-type picture): (*a*) collapse to a black hole, (*b*) explosion from a white hole. An observer's eye is at the top right.

Various authors[52,53] have attempted to invoke white holes as explanations for quasars or other violent astronomical phenomena (sometimes using the name of 'lagging cores'). Recall that in a (classical) collapse to a black hole, the situation starts out with a perfectly normal distribution of matter which follows deterministic evolutionary equations. At a certain stage a trapped surface may form, leading to the presence of an *absolute event horizon* into which particles can fall, but out of which none can escape. After all the available matter has been swallowed, the hole settles

† For this purpose, a 'solution of the Einstein equations' would be a Lorentzian 4-manifold with a *time-orientation* (and perhaps a space-orientation). 'Reversing the time-ordering' amounts to selecting the opposite time-orientation.

Singularities and time-asymmetry. *General relativity an Einstein centenary survey.*
Eds S. W. Hawking and W. Israel, pp. 581–638, Cambridge University Press, 1979

down and remains unchanging until the end of time (possibly at recollapse of the universe). (This ignores the quantum-mechanical effects of the Hawking process,[56] which I shall discuss in a moment.) A white hole, therefore, is created at the beginning of time (i.e. in the big bang) and remains in an essentially unchanging state for an indefinite period. Then it disappears by exploding into a cloud of ordinary matter. During the long quiescent period, the boundary of the white hole is a stationary horizon – the *absolute particle horizon* – into which no particle can fall, but out through which particles may, in principle, be ejected.

There is something that seems rather 'thermodynamically unsatisfactory' (or physically improbable) about this supposed behaviour of a white hole, though it is difficult to pin down what seems wrong in a definitive way. The normal picture of collapse to a black hole seems to be 'satisfactory' as regards one's conventional ideas of classical determinism. Assuming that (strong) cosmic censorship[57–59] holds true, the entire spacetime is determined to the *future* of some 'reasonable' Cauchy hypersurface, on which curvatures are small. But in the case of the white hole, there is no way of specifying such boundary conditions in the past because an initial Cauchy hypersurface has to encounter (or get very close to) the singularity. Put another way, the future behaviour of such a white hole does not, in any *sensible* way, seem to be determined by its past. In particular, the precise moment at which the white hole explodes into ordinary matter seems to be entirely of its own 'choosing', being unpredictable by the use of normal physical laws. Of course one could use future boundary conditions to retrodict the white hole's behaviour, but this (in our entropy-increasing universe) is the thermodynamically unnatural way around. (And, in any case, one can resort to memory as a more effective means of retrodiction!)

Related to this is the fact that while an external observer (using normal retarded light) cannot directly see the singularity in the case of a black hole, he can do so in the case of a white hole (figure 12.5). Since a spacetime singularity is supposed to be a place where the known physical laws break down it is perhaps not surprising, then, that this implies a strong element of indeterminism for the white hole. Causal effects of the singularity can, in this case, influence the outside world.

The presently accepted picture of the physical effects that are expected to accompany a spacetime singularity, is that *particle creation* should take place.[62,63,66] This is the general conclusion of various different investigations into curved-space quantum field theory. However, owing to the incomplete state of this theory, these investigations do not always agree

on the details of their conclusions. As applied to a white hole, two particular schools of thought have arisen. According to Zel'dovich[64] the white hole ought to be completely unstable to this process, evaporating away instantaneously, while Hawking[67] has put forward the ingenious viewpoint that the white hole evaporation ought to be much lower, and indistinguishable in nature from that produced, according to the Hawking process, by a black hole of the same mass – indeed, that a white hole ought itself to be indistinguishable from a black hole! This Hawking viewpoint is, in a number of respects, a very radical one which carries with it some serious difficulties. I shall consider these in a moment. The other viewpoint has the implication that white holes should not physically exist (though, owing to the tentative nature of the particle-creation calculations, this may not carry a great deal of weight; however, cf. also Eardley[65]).

There is another reason for thinking of white holes as antithermodynamic objects (though this reason, too, must be modified if one attempts to adopt the above-mentioned more radical of Hawking's viewpoints). According to the Bekenstein–Hawking formula,[56,69] the surface area A of a *black* hole's absolute event horizon is proportional to the intrinsic entropy S of the hole:

$$S = kAc^3/4\hbar G$$

(k being Boltzmann's constant and G being Newton's gravitational constant). The area principle of classical general relativity[7,70,71] tells us that A is non-decreasing with time in classical processes, and this is compatible with the thermodynamic time's arrow that entropy should be non-decreasing. Now, if a white hole is likewise to be attributed an intrinsic entropy, it is hard to see how the value of this entropy can be other than that given by the Bekenstein–Hawking formula again, but where A now refers to the absolute *particle* horizon. The time-reverse of the area principle then tells us that A is *non-increasing* in classical processes, which is the opposite of the normal thermodynamic time's arrow for entropy. In particular, the value of A will substantially *decrease* whenever the white hole ejects a substantial amount of matter, such as in the final explosion shown in figure 12.5. Thus, this is a strongly anti-thermodynamic behaviour.

It seems that there are two main possibilities to be considered concerning the physics of white holes. One of these is that there is a general principle that rules out their existence (or, at least, that renders them overwhelmingly improbable). The other possibility is contained in the

aforementioned line of argument due to Hawking, which suggests that because of quantum-mechanical effects, black and white holes are to be regarded as physically indistinguishable.[67] I shall discuss, first, Hawking's remarkable idea. Then I shall attempt to indicate why I nevertheless believe that this cannot be the true explanation, and that it is necessary that white holes do *not* physically exist.

Recall, first, the Hawking radiation that is calculated to accompany any black hole. The temperature of the radiation is inversely proportional to the mass of the hole, being of the general order of 10^{-7} K in the case of a black hole of 1 $M_{\odot}$. Of course this temperature is utterly insignificant for stellar-mass holes, but it could be relevant observationally for very tiny holes, if such exist. In an otherwise empty universe, the Hawking radiation would cause the black hole to lose mass, become hotter, radiate more, lose more mass, etc., the whole process accelerating until the hole disappears (presumably) in a final explosion. But for black hole of solar mass (or more) the process would take $>10^{53}$ Hubble times! And, so that the process could even begin, a wait of 10^{7}, or so, Hubble times would be needed to enable the expansion of the universe to reduce the present background radiation to below that of the hole – assuming an indefinitely expanding universe-model!

The absurdity of such figures notwithstanding, it is of some considerable theoretical interest to contemplate, as Hawking has done,[67,72] the state of thermal equilibrium that would be achieved by a black hole in a large container with perfectly reflecting walls. If the container is sufficiently large for a given total mass–energy content (case (*a*)), the black hole will radiate itself away completely (presumably) – after having swallowed whatever other stray matter there had been in the container – to leave, finally, nothing but thermal radiation (with perhaps a few thermalized particles). This final state will be the 'thermal equilibrium' state of maximum entropy (see figure 12.6(*a*)).

If the container is substantially smaller (case (*c*)) – or, alternatively, if the mass–energy content is substantially larger (though still not large enough to collapse the whole container) – the maximum-entropy state will be achieved by a single spherical black hole in thermal equilibrium with its surrounding radiation. Stability is here achieved because, if by a fluctuation the hole radiates a bit too much and consequently heats up, its surroundings heat up even more and cause it to absorb more than it emits and thus to return to its original size; if by a fluctuation it radiates less than it absorbs, its surroundings cool by more than it does and again it returns to equilibrium (figure 12.6(*c*)).

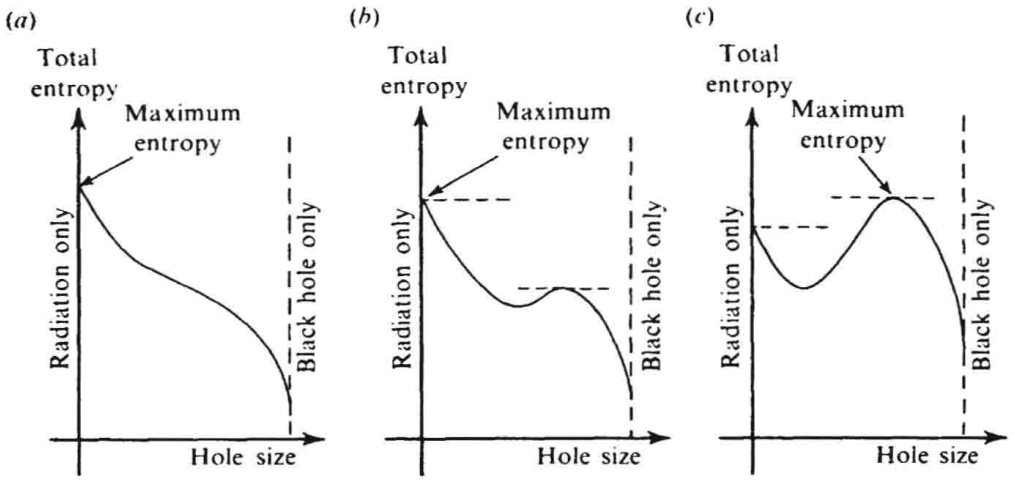


Figure 12.6. Hawking's black hole in a perfectly reflecting container: (*a*) large container, (*b*) intermediate container, (*c*) smallish container.

There is also a situation (case (*b*)) in which the container lies in an intermediate size range, for the given mass–energy content, and where the black hole is still stable, but only represents a *local* entropy maximum, the absolute maximum being a state in which there is only thermal radiation (and perhaps thermalized particles) but no black hole. In this case the black hole can remain in equilibrium with its surrounding radiation for a very long period of time. A large fluctuation would be needed in which a considerable amount of radiation is emitted by the hole, sufficient to get across the low entropy barrier between the two local maxima (figure 12.6(*b*)). With such a large mass loss from the hole, it is able to heat up by an amount greater than its surroundings are able to do; it then loses more mass, heats up more, and, as in case (*a*), radiates itself away completely, to give the required state of thermal radiation (plus occasional thermalized particles).

It should be pointed out that whereas we are dealing with processes that have absurdly long timescales,† these situations have a rather fundamental significance for physics. We are concerned, in fact, with the states of maximum entropy for *all* physical processes. In cases (*a*) and (*b*), the maximum-entropy state is the familiar 'heat death of the universe', but in case (*c*) we have something new: a black hole in thermal equilibrium with radiation. There are, of course, many detailed theoretical difficulties with this setup (e.g. the Brownian motion of the black hole would occasionally send it up against a wall of the container, whereupon

† If we allow (virtual) black holes of down to 10^{-33} cm (i.e. $\sim 10^{-20}$ of an elementary particle radius) then we obtain a picture[73] for which these timescales can be very short.

the container should be destroyed). Such problems will be ignored as irrelevant to the main issues! But also, since the relaxation times are so much greater than the present age of the universe, the interpretations of one's conclusions do need some care. Nevertheless, I feel that there are important insights to be gained here.

To proceed further with Hawking's argument, consider case (*c*). For most of the time the situation remains close to maximum entropy: a black hole with radiation. But occasionally, via an initial large fluctuation in which a considerable energy is emitted by the hole, a sequence like that just considered for case (*b*) will occur, where the black hole evaporates away to give thermal radiation. But then, after a further long wait, enough radiation (again by a fluctuation) collects together in a sufficiently small region for a black hole to form. Provided this hole is large enough, the system settles back into the maximum-entropy state again, where it remains for a very long while.

Cycles like this can also occur in case (*b*) (and even in case (*a*)), but with the difference that most of the time is spent in a state where there is no black hole. In case (*c*), a black hole is present most of the time. Hawking now argues that since the essential physical theories involved are time-symmetric (general relativity, Maxwell theory, neutrinos, possibly electrons, pions, etc., and the general framework of quantum mechanics), the equilibrium states ought to be time-symmetric also. But reversing the time-sense leads to white holes, not to black holes. Thus, Hawking proposes, white holes ought to be physically indistinguishable from black holes!

This identification is not so absurd as one might think at first. The Hawking radiation from the black hole becomes reinterpreted as particle creation near the singularity of the white hole (and hence Hawking proposes a rather slow rate of particle production at the white hole singularity). The swallowing of radiation by the black hole becomes time-reversed Hawking radiation from the white hole. One can, of course, envisage a black hole swallowing a complicated object such as a television set. How can this be thought of as time-reversed Hawking radiation? The argument is that Hawking radiation, being thermal,[67,74] produces all possible configurations with equal probability. It is *possible* to produce a television set as part of the Hawking radiation of a black hole, but such an occurrence is overwhelmingly improbable and would correspond to a large reduction in entropy. A black hole swallowing a television set only seems more 'natural' because we are used to situations in which the entropy is low in the initial state. We can equally well

envisage initial boundary conditions of low entropy for the time-reversed Hawking radiation – and this would be the case for a television set being annihilated as time-reversed Hawking radiation of a white hole.

So far, this all seems quite plausible, and there is even a certain unexpected elegance and economy in the whole scheme. But unfortunately is suffers from two (or perhaps three) very severe drawbacks which, in my opinion, rule it out as a serious possibility.

In the first place, whereas the geometry of the spacetime outside a stationary black hole's horizon is identical to that outside a stationary white hole's horizon, it is definitely *not* so that the exterior geometry of a black hole that forms by standard gravitational collapse and then finally disappears according to the Hawking process is time-symmetric. This time-asymmetry is made particularly apparent by use of conformal diagrams as shown in figure 12.7. A precise distinction between the

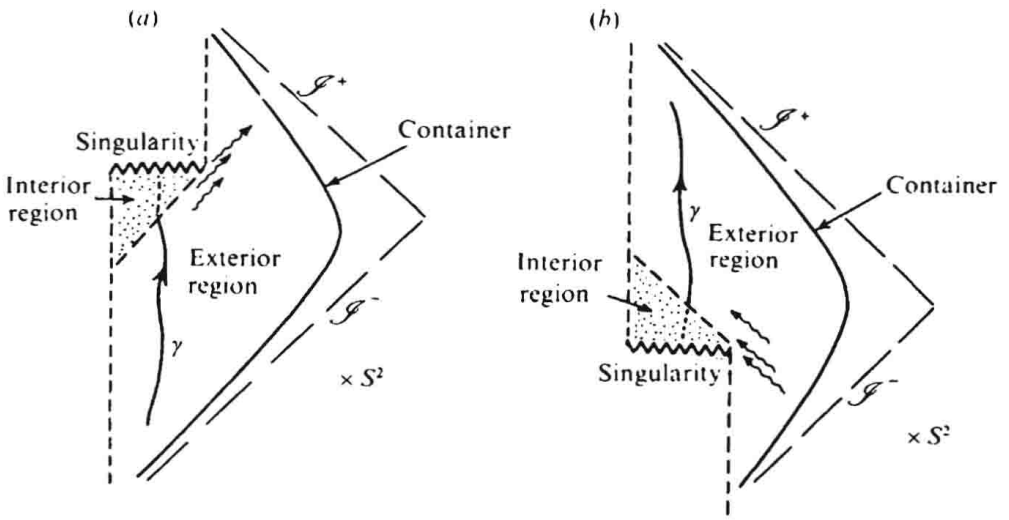


Figure 12.7. Conformal diagrams illustrating time-asymmetry of a transient black hole. (*a*) Classical collapse to a black hole followed by complete Hawking evaporation. (*b*) Hawking condensation to a white hole followed by its classical disappearance.

transient black hole and white hole external geometries can be made in terms of their TIP and TIF structures (cf. section 12.3.2). But intuitively, the distinction should be clear from the presence of timelike curves γ which, in the case of the black hole, 'leave' the exterior geometry to 'enter' the hole; and the other way around in the case of the white hole. The reason for this distinction is simply that the process of *classical* collapse is not the time-reverse of the *quantum* Hawking process. This should not really surprise us since each relies on quite different physical

theories (classical general relativity as opposed to quantum field theory on a fixed curved-space background).

A point of view adopted by Hawking which might avoid this difficulty is to regard the spacetime geometry as being somewhat observer-dependent. Thus, as soon as quantum mechanics and curved-space geometry have become essentially intertwined, so this viewpoint would maintain, one cannot consistently talk about a classically objective spacetime manifold. An observer who falls into a white hole to be evaporated away as its time-reversed Hawking radiation would, accordingly, believe the geometry to be, instead, that of a black hole whose horizon he crosses and inside which he awaits his 'classical' fate of final destruction by excessive tidal forces.

I have to say that I find this picture almost as hard to accept as those according to which the entropy starts decreasing when the observer's universe collapses about him. If one is considering black or white holes whose radius is of the order of the Planck length ($\sim 10^{-33}$ cm) – or even, possibly, of the order of an elementary particle size ($\sim 10^{-13}$ cm) – then such indeterminancy in the geometry might be acceptable. But for a black hole of solar mass (or more) this would entail a very radical change in our views about geometry, a change which would drastically affect almost any other application of general relativity to astrophysical phenomena. It is true that in section 12.2.5 I have briefly entertained the possibility of a world-view which allows for an element of 'observer-dependence' in the geometry. But I have as yet seen no way to relate such a view to the kind of indeterminacy in classical geometry that the physical identification of black holes with white holes seems to lead to.

But there are also other objections to attempting to regard classical gravitational collapse as being effectively the time-reverse of a quantum-mechanical particle creation process. One of these refers not so much to the attempted identification of the Hawking process with the time-reverse of the classical swallowing of matter by a black hole, but with the attempt to identify either of these processes with the phenomenon of particle production at regions of large spacetime curvature. Such a further identification *seems* to be an integral part of the time-symmetric view that I have been discussing, though it may well not really be what is intended. I am referring to the picture of Hawking radiation by a black hole as being alternatively regarded as a process of particle production near the singularity of a white hole. If, indeed, it can be so regarded, then this is not the 'normal' process of particle production at regions of large curvature that has many times been discussed in the literature.[62] For in

that process, particles are always produced in *pairs*: baryon with antibaryon, lepton with antilepton; positively charged particle with negatively charged particle. But the Hawking process is explicitly not of this form, as its thermal nature (for particles escaping to infinity) implies.[67,74,75]

The contrast is even more blatant if we try to relate this pairwise particle production near white hole singularities to the time-reverse of the destruction of matter at a black hole's singularity. For there are no constraints whatever on the type of matter that a black hole can classically absorb. And if strong cosmic censorship is accepted in classical processes, it seems that even individual charged particles have then to be separately destroyed at the singularity. (This point will be amplified in section 12.3.2.) There is no suggestion that the particles must somehow contrive to sort themselves into particle–antiparticle pairs before they encounter the singularity. The difficulty here is, perhaps, not so much directly to do with Hawking's identification of black holes with white holes, but with the whole idea of hoping to deal with the matter destruction–creation process in terms of known physics. Thus I maintain that, whereas it *may* be that matter creation at the big bang can be treated in terms of known (or at least partially understood) particle-creation processes, this seems *not* to be true of the destruction processes at black hole singularities – nor, if Hawking is right, of the creation processes at white hole singularities. Thus, the Hawking view would seem to lead to direct conflict with the often expressed hope that particle creation at the big bang can be understood in terms of processes of particle production by spacetime curvature. This relates to the question of whether time-symmetric physics can be maintained at spacetime singularities. This is a key issue that I shall discuss in more detail in section 12.3.

More directly related to Hawking's proposal is a difficulty which arises if we examine in detail the cycles whereby a solar-mass black hole (say), in stable equilibrium with its surroundings, in our perfectly reflecting container, may disappear and reform, owing to fluctuations, in cases (*b*) and (*c*) just discussed. What, in fact, is the most probable way for the black hole to evaporate completely? It might, of course, simply throw out its entire mass in one gigantic fluctuation. But the random Hawking process would achieve this only absurdly infrequently. Overwhelmingly *less* absurdly infrequent would be the emission, in one huge fluctuation, of that *fraction* of the hole's mass needed to raise its Bekenstein–Hawking temperature above that to which its surroundings would be consequently raised. From there on, the evaporation would proceed

'normally' needing no further improbable occurrence. But consider the final explosion according to which the hole 'normally' disappears. Electrons and positrons have been appearing, followed by pions; then, at the last moment, a whole host of unstable particles is produced which undergo complicated decays. Finally one expects many protons and antiprotons separately escaping the annihilation point. Only much later, by chance encounters as they move randomly about the container, would one expect the protons and antiprotons gradually to annihilate one another (or possibly occasionally to decay, themselves, by a Pati–Salam-type process[76] into, say, positrons and into electrons which would then mostly annihilate one another).

What, now, must we regard as the most probable way in which a black hole forms again in the container, to reach another point of stable equilibrium with its surroundings? Surely it is *not* the time-reverse of the above, according to which the protons and antiprotons must first (with long preparation) contrive to form themselves (quite unnecessarily) out of the background radiation before aiming themselves with immense accuracy at a tiny point, only to indulge in some (again unnecessary) highly contrived particle physics whereby they meet up with carefully aimed γ-rays, etc., to form various unstable particles, etc., etc.; and then (also with choreography of the utmost precision) other particles (pions, then electrons and positrons) must aim themselves inwards, having first formed themselves out of the background at the right moment and in the correct proportions. Only later does the background radiation itself fall inwards to form the bulk of the mass needed to form the hole.

My point is not that this curious beginning is necessarily the most improbable part of the process. I can imagine that it may well not be. But it is unnecessary. The essential part of the hole formation of a *black* hole would occur when the radiation itself collects together in a sufficiently small region to undergo what is, in effect, a standard gravitational collapse. In fact, it would seem that the tiny core that has been formed with such elaborate preparations ought somewhat to *inhibit* the subsequent inward collapse, owing to its excessive temperature!

So what has gone wrong? What time-asymmetric physics has been smuggled into the description of the 'most probable' mode of disappearance of the hole, that it should so disagree with the time-reverse of its 'most probable' mode of reappearance? Possibly none, if white holes, in principle, exist, and are simply *different* objects from black holes. While the above elaborate preparations are not necessary for the formation of a black hole, they could be for a white hole. After all, one has to contrive

some way of producing the white hole's initial singularity which, as is evident from figure 12.7, has a quite different structure from that of a black hole. It might be that the production of such a singularity is an extraordinarily delicate process, requiring particles of just the right kind and in an essentially right order to be aimed with high energy and with extraordinary accuracy. There is an additional seeming difficulty, however, in that one also has to conjure up a region of spacetime (namely that inside the horizon) which does not lie in the domain of dependence of some initial Cauchy hypersurface drawn before the white hole appears (figure 12.7(*b*)). Of course, the time-reverse of this problem also occurs for the (Hawking) disappearance of a black hole, but one is not in the habit of trying to retrodict from Cauchy hypersurfaces, so this seems less worrisome! An additional difficulty, if white holes are allowed, is that we now encounter the problem, mentioned earlier, of trying to predict what the white hole is going to emit, and when. As I have indicated, *if* time-symmetric physics holds at singularities, then 'normal' ideas of particle creation due to curvature will not do. Hawking's concept of 'randomicity'[67] might be nearer the mark, but it is unfortunately too vague to enable any calculations to be made – now that the essential guiding idea of an identification between black holes and white holes has been removed.

I find the picture of an equilibrium involving the occasional production of such white holes a very unpleasant one. And what other monstrous zebroid combinations of black and white might also have to be contemplated? I feel that such things have nothing really to do with physics (at least on the macroscopic scale). The only reason that we have had to consider white holes *at all* is in order to save time-symmetry! The consequent unpleasantness and unpredictability seems a high price to pay for something that is *not even true* of our universe on a large scale.

One of the consequences of the hypothesis that I shall set forth in the next section is that it rules out the white hole's singularity as an unacceptable boundary condition. The hypothesis is time-asymmetric, but this is necessary in order to explain the other arrows of time. When we add this hypothesis to the discussion of equilibrium within the perfectly reflecting container, we see explicitly what time-asymmetric physics has been 'smuggled' in. For the hypothesis is designed not to constrain the behaviour of black holes in any way, but it forbids white holes and therefore renders irrelevant the extraordinary scenario that we seem to need in order to produce one!

Singularities and time-asymmetry. *General relativity an Einstein centenary survey.*
Eds S. W. Hawking and W. Israel, pp. 581–638, Cambridge University Press, 1979

I hope the reader will forgive me for having discussed white holes at such length only to end by claiming that they do not exist! But hypothetical situations can often lead to important understandings, especially when they border on the paradoxical, as seems to be the case here.

12.3 Singularities: the key?

What is the upshot of the discussion so far? According to sections 12.2.3, 12.2.4 and 12.2.5, the arrows of entropy and retarded radiation, and *possibly* of psychological time, can all be explained if a reason is found for the initial state of the universe (big bang singularity) to be of comparatively low entropy and for the final state to be of high entropy. According to section 12.2.6, some low-entropy assumption *does* need to be *imposed* on the big bang; that is, the mere fact that the universe expands away from a singularity is in no way sufficient. And according to section 12.2.7, we need some assumption on initial singularities that rules out those which would lie at the centres of white holes. On the other hand, the discussions in sections 12.2.1 and 12.2.2 were inconclusive, and I shall need to return to them briefly at the end.

But what is it in the nature of the big bang that is of 'low entropy'? At first sight, it would seem that the knowledge we have of the big bang points in the opposite direction. The matter (including radiation) in the early stages appears to have been completely thermalized (at least so far as this is possible, compatibly with the expansion). If this had not been so, one would not get correct answers for the helium abundance, etc.[77,78] And it is often remarked upon that the 'entropy per baryon' (i.e. the ratio of photons to baryons) in the universe has the 'high' value of $\sim 10^9$. Ignoring the contribution to the entropy due to black holes, this value has remained roughly constant since the very early stages, and then represents easily the major contribution to the entropy of the universe – despite all the 'interesting' processes going on in the world, so important to our life here on Earth, that depend upon 'small' further taking up of entropy by stars like our Sun. The answer to this apparent paradox – that the big bang thus *seems* to represent a state of *high* entropy – lies in the unusual nature of gravitational entropy. This I next discuss, and then show how this relates to the structure of singularities.

12.3.1 Gravitational entropy

It has been pointed out by many authors[79] that gravity behaves in a somewhat anomalous way with regard to entropy. This is true just as

much for Newtonian theory as for general relativity. (In fact, the situation is rather worse for Newtonian theory.) Thus, in many circumstances in which gravity is involved, a system may behave as though it has a negative specific heat. This is directly true in the case of a black hole emitting Hawking radiation, since the more it emits, the hotter it gets. But even in such familiar situations as a satellite in orbit about the Earth, we observe a phenomenon of this kind. For dissipation (in the form of frictional effects in the atmosphere) will cause the satellite to speed up, rather than slow down, i.e. cause the kinetic energy to increase.

This is essentially an effect of the universally attractive nature of the gravitational interaction. As a gravitating system 'relaxes' more and more, velocities increase and the sources clump together – instead of uniformly spreading throughout space in a more familiar high-entropy arrangement. With other types of force, their attractive aspects tend to saturate (such as with a system bound electromagnetically), but this is not the case with gravity. Only non-gravitational forces can prevent parts of a gravitationally bound system from collapsing further inwards as the system relaxes. Kinetic energy itself can halt collapse only temporarily. In the absence of significant non-gravitational forces, when dissipative effects come further into play, clumping becomes more and more marked as the entropy increases. Finally, maximum entropy is achieved with collapse to a black hole – and this leads us back into the discussion of section 12.2.7.

Consider a universe that expands from a 'big bang' singularity and then recollapses to an all-embracing final singularity. As was argued in section 12.2.6, the entropy in the late stages ought to be much higher than the entropy in the early stages. How does this increase in entropy manifest itself? In what way does the high entropy of the final singularity distinguish it from the big bang, with its comparatively low entropy? We may suppose that, as is apparently the case with the actual universe, the entropy in the initial *matter* is high. The kinetic energy of the big bang, also, is easily sufficient (at least on average) to overcome the attraction due to gravity, and the universe expands. But then, relentlessly, gravity begins to win out. The precise moment at which it does so, locally, depends upon the degree of irregularity already present, and probably on various other unknown factors. Then clumping occurs, resulting in clusters of galaxies, galaxies themselves, globular clusters, ordinary stars, planets, white dwarfs, neutron stars, black holes, etc. The elaborate and interesting structures that we are familiar with all owe their existence to

this clumping, whereby the gravitational potential energy begins to be taken up and the entropy can consequently begin to rise above the *apparently* very high value that the system had initially. This clumping must be expected to increase; more black holes are formed; smallish black holes swallow material and congeal with each other to form bigger ones. This process accelerates in the final stages of recollapse when the average density becomes very large again, and one must expect a very irregular and clumpy final state.

There is a slight technical difficulty in that the concept of a black hole is normally only defined for asymptotically flat (or otherwise open) spacetimes. This difficulty could affect the discussion of the final stages of collapse when black holes begin to congeal with one another, and with the final all-embracing singularity of recollapse. But I am not really concerned with the location of the black holes' event horizons, and it is only in precisely defining these that the aforementioned difficulty arises. A black hole that is formed early in the universe's history has a singularity that is reached at early proper times for observers who encounter it;[57] for holes that are formed later, they can be reached at later proper times. On the basis of strong cosmic censorship (cf. section 12.3.2), one expects all these singularities eventually to link up with the final singularity of recollapse.[57] I do not require that the singularities of black holes be, in any clear-cut way, distinguishable from each other or from the final singularity of recollapse. The point is merely that the gravitational clumping which is characteristic of a state of high gravitational entropy should manifest itself in a very complicated structure for the final singularity (or singularities).

The picture is not altogether dissimilar for a universe that continues to expand indefinitely away from its big bang. We still expect local clumping, and (provided that the initial density is not altogether too low or too uniform for galaxies to form at all) a certain number of black holes should arise. For the regions inside these black holes, the situation is not essentially different from that inside a collapsing universe (as was remarked upon in section 12.2.6), so we expect to find, inside each hole, a very complicated singularity corresponding to a very high gravitational entropy. For those regions not inside black holes there will still be certain localized portions, such as rocks, planets, black dwarfs, or neutron stars, which represent a certain ultimate raising of the entropy level owing to gravitational clumping, but the gain in gravitational entropy will be relatively modest, though sufficient, apparently, for all that we need for life here on Earth.

I have been emphasizing a qualitative relation between gravitational clumping and an entropy increase due to the taking up of gravitational potential energy. In terms of spacetime curvature, the absence of clumping corresponds, very roughly, to the absence of Weyl conformal curvature (since absence of clumping implies spatial-isotropy, and hence no gravitational principal null-directions).[45] When clumping takes place, each clump is surrounded by a region of nonzero Weyl curvature. As the clumping gets more pronounced owing to gravitational contraction, new regions of empty space appear with Weyl curvature of greatly increased magnitude. Finally, when gravitational collapse takes place and a black hole forms, the Weyl curvature in the interior region is larger still and diverges to infinity at the singularity.

At least, that is the picture presented in spherically symmetric collapse, the magnitude of the Weyl curvature diverging as the inverse cube of the distance from the centre. But there are various reasons for believing that in generic collapse, also, the Weyl curvature should diverge to infinity at the singularity, and (at most places near the singularity) should dominate completely over the Ricci curvature.

This can be seen explicitly in the details of the Belinskii–Khalatnikov–Lifshitz analysis.[80] Moreover, one can also infer it on crude qualitative grounds. In the exact Friedmann models, it is true, the Ricci tensor dominates, the Weyl tensor being zero throughout. In these cases, as a matter world-line is followed into the singularity, it is approached isotropically by the neighbouring matter world-lines, so we have simultaneous convergence in three mutually perpendicular directions orthogonal to the world-line. In the case of spherically symmetrical collapse to a black hole, on the other hand, if we envisage some further matter falling symmetrically into the central singularity, it will normally converge in towards a given matter world-line only in *two* mutually perpendicular directions orthogonal to the world-line (and diverge in the third). This is the situation of the Kantowski–Sachs[81] cosmological model, giving a so-called 'cigar'-type singularity.[7] If r is the usual Schwarzschild coordinate, the volume gets reduced like $r^{3/2}$ near the singularity, so the densities are $\sim r^{-3/2}$. Thus, for a typical Ricci tensor component, $\Phi \sim r^{-3/2}$. However, in general, for a typical Weyl tensor component, $\Psi \sim r^{-3}$, showing that the Weyl tensor dominates near the singularity in these situations. Also, in the 'pancake' type of singularity, where there is convergence in only *one* direction orthogonal to a matter world-line, we again expect the Weyl tensor to dominate with $\Phi \sim r^{-1}$ and $\Psi \sim r^{-2}$ in this case.

Singularities and time-asymmetry. *General relativity an Einstein centenary survey.* Eds S. W. Hawking and W. Israel, pp. 581–638, Cambridge University Press, 1979

Now the Friedmann type of situation, with simultaneous convergence of all matter from all directions at once, would seem to be a very special setup. If there is somewhat less convergence in one direction than in the other two, then a cigar-type configuration seems more probable very close to the singularity, while a pancake-type appears to result when the main convergence is only in one direction. Moreover, with a generic setup, a considerable amount of oscillation seems probable.[80] An oscillating Weyl curvature of frequency ν and complex amplitude Ψ, supplies an *effective* additional 'gravitational-energy' contribution to the Ricci tensor[61] of magnitude $\sim|\Psi|^2\nu^{-2}$. If ν becomes very large so that many oscillations occur before the singularity is reached, then[49] $\nu^2 \gg \Phi^{-1}$, where Φ is a typical Ricci tensor component. Thus if, as seems reasonable in general, the 'energy content' of Ψ is to be comparable with Φ as the singularity is approached, we have $|\Psi|^2\nu^{-2} \sim \Phi$, so $|\Psi| \gg \Phi$. These considerations are very rough, it is true, but they seem to concur with more detailed analysis[80] which indicates that in generic behaviour near singularities the contributions due to matter can be ignored to a first approximation and the solution treated as though it were a vacuum, i.e. that the Weyl part of the curvature dominates over the Ricci part.

The indications are, then, that a high-entropy singularity should involve a very large Weyl curvature, unlike the situation of the singularity in the Friedmann dust-filled universe or any other models of the Robertson–Walker class. At the time of writing, however, no clear-cut integral formula (say) which could be regarded as giving mathematical expression to this suggested relation between Weyl curvature and gravitational entropy has come to light. Some clues as to the nature of such a formula (if such exists at all) may be obtained, firstly, from the Bekenstein–Hawking formula for the entropy of a black hole and, secondly, from the expression for the particle number operator for a linear spin-2 massless quantized free field – since an estimate of the 'number of gravitons' in a gravitational field could be taken as a measure of its entropy.† Thus this entropy measures the number of *quantum* states that contribute to a given classical geometry.

There is one final point that should be mentioned in connection with the question of the entropy in the gravitational field. It was pointed out some time ago by Tolman[84] that a model universe containing matter that

† This point of view does not seem to agree with that of Gibbons and Hawking,[82] who apparently regard the gravitational entropy as being zero when black holes are absent. But an estimate of 'photon number' in a classical electromagnetic field gives a measure of its entropy[83] (without black holes). Gravity is presumably similar.

appeared to be in thermal equilibrium in its early stages can lead to a situation in which the matter gets out of equilibrium as the universe expands (a specific example of matter illustrating such behaviour being a diatomic gas which is capable of dissociating into its elements and recombining). Then, if such a model represents an expanding and recollapsing universe, the state of the matter during recollapse would differ from the corresponding state during expansion, where we make the correspondence at equal values of the universe radius R (or comoving radius R). In fact, the matter, during recollapse, would have acquired some energy out of the global geometry of the universe, the resulting difference in geometry showing up in the fact that $\dot{R}^2$ is greater, for given R, at recollapse than it is during the expansion. So the entropy of the system as a *whole* increases with time even though the matter *itself* is in thermal equilibrium during an initial stage of the expansion. There is, in fact, a contribution to the entropy from R (and $\dot{R}$), which must be regarded as a dynamical variable in the model. (This arises because of the phenomenon of *bulk viscosity*.[85])

One can view what is involved here as basically a transfer of potential energy from the global structure of the universe (gravitational potential energy) into the local energy of the matter, though there are well-known difficulties about defining energy in a precise way for models of this kind. But these difficulties should not concern us unduly here, since it is actually the entropy rather than the energy that is really relevant, and entropy has much more to do with probabilities and coarse-graining than it has to do with any particular definition of energy. In the example given by Tolman there is no state of maximum entropy, either achieved in any one specific model or throughout all models of this type. By choosing the value of R at maximum expansion to be sufficiently large (for fixed matter content), the total entropy can be made as large as we please. Tolman envisaged successive cycles of an 'oscillating' universe with gradually increasing maximum values of R. However, it is hard for us to maintain such a world-view now, because the singularity theorems[7,8] tell us that the universe cannot achieve an effective 'bounce' at minimum radius without violating the known† laws of physics.

From my own point of view, the situation envisaged by Tolman may be regarded as one aspect of the question of how the structure of the universe as a whole contributes to the entropy. It apparently concerns a somewhat different aspect of this question than does gravitational

† I am counting quantum gravity as 'unknown' whether or not it helps with the singularity problem!

clumping, since the Weyl tensor is everywhere zero in Tolman's models. It is clear that this has also to be understood in detail if we are to perceive, fully, the role of gravitational entropy. Nevertheless, it appears that the entropy available in Tolman's type of situation is relatively insignificant[78] compared with that which can be obtained – and, indeed, *is* obtained – by gravitational clumping (cf. section 12.3.3).

The key question must ultimately concern the structure of the singularities. These singularities, in any case, provide the boundary conditions for the various cycles in Tolman's 'oscillating' universe. Moreover, as we shall see in a moment, if strong cosmic censorship holds true, the presence of irregularities should not alter the all-embracing nature of these singularities in the case of an expanding and recollapsing universe.

12.3.2 Cosmic censorship†

Though it is by no means essential, for the viewpoint that I am proposing, to suppose that naked singularities cannot occur, such an assumption of 'cosmic censorship' does nevertheless greatly simplify and clarify the discussion. It has been my stated intention to adopt basically conventional attitudes on most issues, so I should not be out of line here, also, were I simply to align myself with what appears to be a majority view and (at least for purposes of argument) adopt a suitable assumption of cosmic censorship. But in the following pages I shall also give some independent justification of this view.

A preliminary remark is required before considering the details of this, however. In the Hawking process of black hole evaporation, the (supposed) final disappearance of the hole produces, momentarily, a naked singularity. This is not normally considered to be a violation of cosmic censorship because the Hawking process is a quantum-mechanical process, whereas cosmic censorship is taken normally to refer only to classical general relativity. (In the words of Hawking,[68] cosmic censorship is 'transcended' rather than violated!) Nevertheless, the presence of actual naked singularities of this kind in the geometry of the world would make a certain difference to the discussion. But the difference seems unlikely to be of any great relevance to the problem at hand. The black holes that we have any clear reason to believe actually exist in the universe are all of the order of a solar mass or more – and we have seen

† Parts of this section are considerably more technical than the others, and it can be omitted without seriously affecting the thread of argument.

that their lifetimes are greater than 10^{53} Hubble times, so their final naked singularities (!) can be safely ignored. Moreover, an implication of the viewpoint I shall set out in section 12.3.3 seems to be that mini-holes are unlikely to exist, these being the only black holes (say of mass 10^{20} g or less) whose final naked singularity could occur early enough to be of any remote relevance to the discussion. But since, in any case, such holes would be of no greater than atomic dimension, they could be 'smoothed over' and need not be considered as constituting a significant part of the classical geometry.

It seems, then, that a discussion of cosmic censorship entirely within classical general relativity should be perfectly adequate for our purposes. So what form of statement should we adopt? The usual formulation involves some assertion such as:

A system which evolves, according to classical general relativity with reasonable equations of state, from generic non-singular initial data on a suitable Cauchy hypersurface, does not develop any spacetime singularity which is visible from infinity.

Something of this kind, forbidding naked singularities, appears to be required in order that the usual general discussion of black holes can be carried through (e.g. the area-increase principle, the merging of two black holes necessarily forming a third, the general macrostability of black holes – indeed the very physical existence of black holes at all, rather than something worse, in a generic collapse[51]). The statement is vague in several respects, and a considerable increase in precision would be required in order to obtain something capable of clear mathematical proof, or disproof.

But it is probably not too helpful simply to make the various conditions more precise in some way, without having a deeper idea of what is likely to be involved. For example, it seems to me to be quite unreasonable to suppose that the physics in a comparatively local region of spacetime should really 'care' whether a light ray setting out from a singularity should ultimately escape to 'infinity' or not. To put things another way, some observer (timelike world-line) might intercept the light ray and see the singularity as 'naked', though he be not actually situated at infinity (and no actual observer would be so situated in any case). The observer might be close by the singularity and possibly himself trapped, e.g. inside the usual black hole of figure 12.5(*a*). The unpredictability entailed by the presence of naked singularities which is so abhorrent to many people

would be present just as much for this local observer – observing a 'locally naked' singularity – as for an observer at infinity.

It seems to me to be comparatively unimportant whether the observer himself can escape to infinity. Classical general relativity is a scale-independent theory, so if locally naked singularities occur on a very tiny scale, they should also, in principle, occur on a very large scale in which a 'trapped' observer could have days or even years to ponder upon the implications of the uncertainties introduced by his observation of such a singularity (compare the analogous discussion in section 12.2.6 of the astronaut inside a large black hole). Indeed, for inhabitants of recollapsing closed universes (as possibly we ourselves are) there is no 'infinity', so the question of being locally 'trapped' is one of degree rather than principle.

It would seem, therefore, that if cosmic censorship is a principle of Nature, it should be formulated in such a way as to preclude such *locally* naked singularities.[57,58,86] This viewpoint gains some support from, first of all, the standard picture of spherically symmetrical collapse inside a black hole as shown in figure 12.5(a). An observer who falls inside the hole cannot, in fact, see the singularity at all until he encounters it. This is perhaps clearer if we use a standard conformal diagram (with null-cones sloping at 45°) to depict the situation, as in figure 12.8, since then the spacelike nature of the singularity is brought out clearly and shows that it is *not* locally naked in the sense described above.

Secondly, there are some reasons for believing that generic perturbations away from spherical symmetry will not change the spacelike nature of the singularity (whose continued existence, in the perturbed case, is guaranteed by the singularity theorems[6–8]). The situation is slightly complicated, however, because the Schwarzschild–Kruskal singularity of figure 12.8 is, in fact, unstable. The introduction of a minute amount of angular momentum into the black hole to give a Kerr solution (with $a \ll m$) will actually change the singularity structure completely, and it ceases to be spacelike. Only when a further perturbation of a generic nature is made, may a spacelike structure of the singularity be expected to be restored.

It is somewhat easier to examine this behaviour if we consider adding charge rather than angular momentum, so that we get the Reissner–Nordstrom solution instead of the Kerr solution. The corresponding conformal diagram is shown in figure 12.9(a), and, indeed, it is evident that the singularity *is* now locally naked in the sense described above, since the observer whose world-line is γ can see the singularity. With a

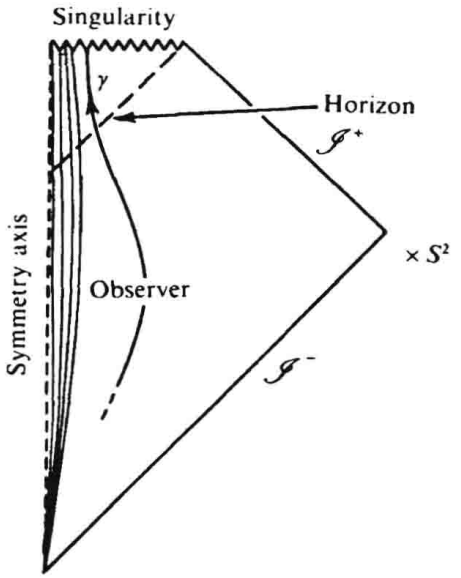


Figure 12.8. Conformal diagram of spherically symmetric collapse (with asymptotic flatness) illustrating the spacelike nature of the singularity.

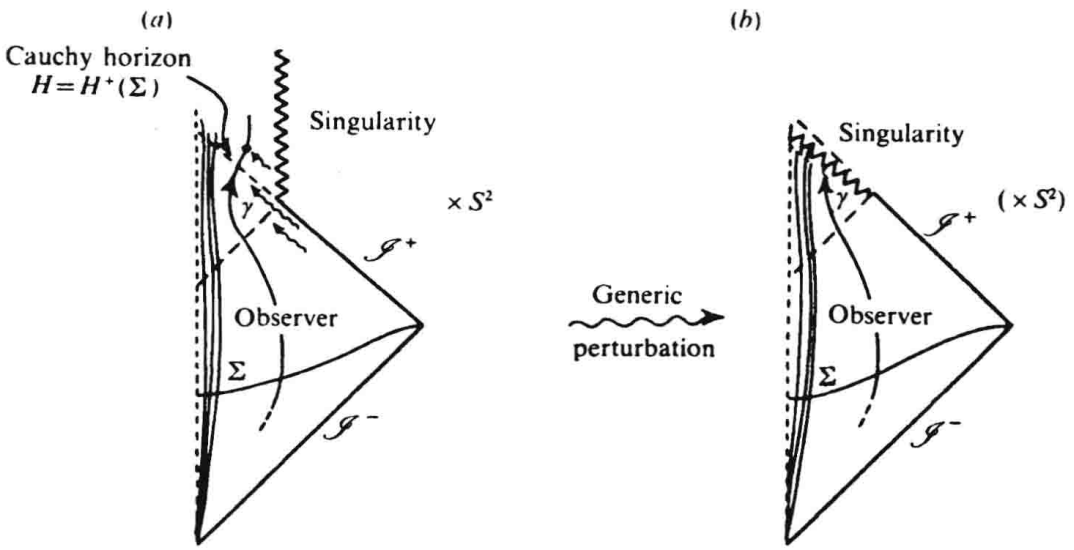


Figure 12.9. Conformal diagrams illustrating collapse to a black hole with a small charge: (*a*) spherically symmetric, (*b*) generically perturbed.

further perturbation of a generic nature one expects the situation more to resemble that of figure 12.9(*b*) in which the singularity is spacelike again (or perhaps null). The reason is that in the situation of figure 12.9(*a*) there is a null-hypersurface, H, in the spacetime, which is the *Cauchy horizon* of an (edgeless) spacelike hypersurface Σ extending all the way out to spatial infinity. (For terminology and notation, see reference 49.) An

observer γ who crosses $H = H^+(\Sigma)$ will see, as he does so, the entire future history of the outside world flash by in an instant. If the data on Σ are perturbed in some mild way out near infinity, then this will lead to a drastic alteration of the geometry in the neighbourhood of H. This is because signals from the infinite regions of Σ will be blueshifted at H by an infinite amount. Indeed, the analysis of weak-field perturbations (or test fields)[87,88] explicitly indicates that these perturbations diverge along H. With the full nonlinear coupling, this would presumably give a curvature singularity in place of H. Furthermore, it *could* well be that nonlinear effects actually result in a spacelike rather than just a null-singularity, because the effects of large curvature might proliferate and reinforce one another further and further up along the singularity (cf. figure 12.9(b)).

In order to make precise the notion of a spacelike (or null) spacetime singularity it will be convenient to recall the concept of an *ideal point*[89] of a spacetime $\mathcal{M}$, defined in terms of the TIPs or TIFs (terminal indecomposable past-sets or future-sets) of $\mathcal{M}$. The ideal points may be thought of as some 'extra points' adjoined to the manifold $\mathcal{M}$, either 'singularities' or 'points at infinity', for which the timelike curves in $\mathcal{M}$ that are future-endless in $\mathcal{M}$ acquire future ideal endpoints (via TIPs), and those that are past-endless in $\mathcal{M}$ acquire past ideal endpoints (via TIFs).

For simplicity, assume that $\mathcal{M}$ is strongly causal. Let γ and γ' be two future-endless timelike curves in $\mathcal{M}$. Then γ and γ' have the same future ideal endpoint if and only if they have the same pasts, which, in standard notation, is written $I^-[\gamma] = I^-[\gamma']$. The TIPs of $\mathcal{M}$ are, in fact, the sets of the form $I^-[\gamma]$, with γ future-endless and timelike, and may be 'identified' with the future ideal points. Similarly, past-endless timelike curves η and η' have the same ideal past endpoints whenever their futures are the same: $I^+[\eta] = I^+[\eta']$, these sets being the TIFs of $\mathcal{M}$. (See figure 12.10). In each case, the TIP or TIF is said to be *generated* by the timelike curve in question. A simple criterion[57,58] that may be used to distinguish those TIPs representing points at infinity from those representing singular points is to define a TIP as an *∞-TIP* if it is generated by some timelike curve of infinite proper length into the future, and as a *singular TIP* if it is generated by no such curve. The *∞-TIFs* and *singular TIFs* are similarly defined. (One may also choose to call some of the ∞-TIPs and ∞-TIFs singular, in some appropriate sense, but I shall not bother with this in detail here.)

Next, a locally *naked* singularity can be defined as either a singular TIP contained in the past $I^-(q)$ of some point q of $\mathcal{M}$, or as a singular TIF

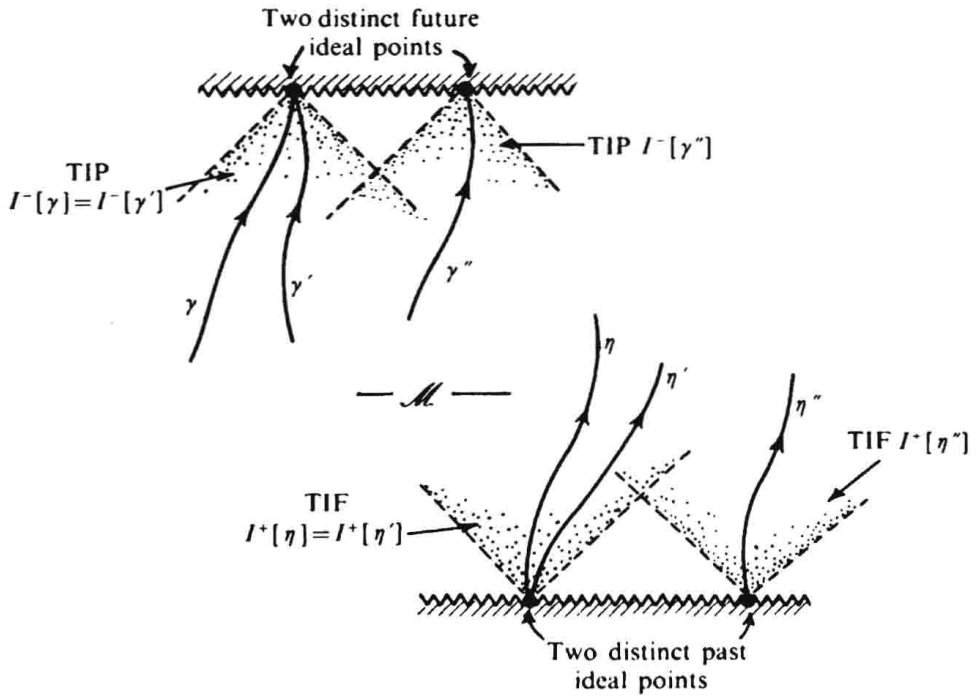


Figure 12.10. TIPs and TIFs defining ideal points of $\mathscr{M}$.

contained in the future $I^+(p)$ of some point p of $\mathscr{M}$ (figure 12.11). The upshot of such a definition is that in each case there is a timelike curve ζ (observer's world-line) from a point p to a point q, where the singularity lies to the future of p and to the past of q. (Take p in the TIP, in the first case, and q in the TIF, in the second.) The significance of this is that not

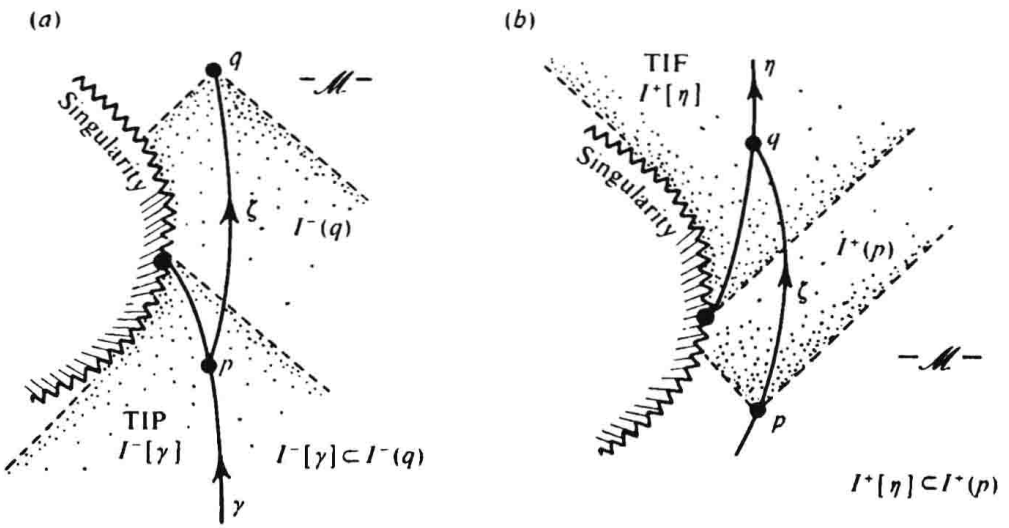


Figure 12.11. A locally naked singularity, lying to the future of p and to the past of q: (*a*) TIP definition, (*b*) TIF definition.

only can some observer (namely ζ) see the singularity (namely from q) but that at some earlier stage of his existence (namely at p) the singularity has yet to be produced. Thus, the usual all-embracing big bang does *not* qualify as locally naked, since no observer existed before it was produced.

One may say that $\mathscr{M}$ accords with a form of cosmic censorship if such locally naked singularities (as illustrated in figures 12.11(a) and (b)) do not occur.[57] However, it is convenient to go somewhat further than this and to exclude 'naked points at infinity' also, by dropping the condition that the TIP or TIF involved in figure 12.11 is necessarily a *singular* TIP or TIF. Indeed, it could be argued that such a naked point at infinity introduces as much indeterminacy into the future behaviour of a universe-model as does a naked singularity. But, in fact, naked points at infinity seem unlikely to arise unless they are also, in some appropriate sense, *singular* points (though still defined by ∞-TIPs or ∞-TIFs). The reason is that for a smooth conformal infinity $\mathscr{I}$, naked points at infinity can occur only if $\mathscr{I}$ is (in some places at least) timelike. And, with reasonable matter density, $\mathscr{I}$ can be timelike only if $\lambda < 0$, where λ is the cosmological constant.[90] However, the usual (Friedmann-type) $\lambda < 0$ models expand from, and recontract to, all-embracing singularities,[91] and hence possess no ∞-TIPs or ∞-TIFs, so (non-singular) naked points at infinity are not to be expected even in these cases.

Whether the TIPs or TIFs refer to singular points or to points at infinity, the situation of figure 12.11 is what characterizes ideal points as belonging to a boundary to $\mathscr{M}$ which is, in a sense, *timelike*. For if we extend ζ indefinitely into the future, in figure 12.11(a), to becomes a future-endless timelike curve ζ' we have

$$I^-[\gamma] \subset I^-(q), \quad q \in I^-[\zeta']$$

which is the condition[89] for the TIP $I^-[\gamma]$ to lie to the chronological (i.e. timelike) past of the TIP $I^-[\zeta']$. (The weaker condition $I^-[\gamma] \subset I^-[\zeta']$ obtains if $I^-[\gamma]$ lies to the *causal* past[89] of $I^-[\zeta']$.) Similarly, figure 12.11(b) gives us

$$I^+[\eta] \subset I^+(p), \quad p \in I^+[\zeta'']$$

with ζ'' past-endless, characterizing the TIF $I^+[\eta]$ as lying to the chronological future of the TIF $I^+[\zeta'']$. Thus, to rule out figure 12.11(a) configurations is to say that the future ideal points constitute an *achronal*[49] (i.e. roughly, spacelike or null) future boundary to $\mathscr{M}$, while to rule out those of figure 12.11(b) is to say that the past ideal points constitute an *achronal* past boundary to $\mathscr{M}$.

What is more, ruling out *either one* of these configurations throughout $\mathcal{M}$ is equivalent to ruling out the other, since each condition turns out to be equivalent[57] to the time-symmetric condition that $\mathcal{M}$ be *globally hyperbolic.*[7,49]

The proof of this fact is quite simple and is worth outlining here since it has not been spelled out in the standard literature. First, a globally hyperbolic spacetime is one with the property that for any two of its points p, q, the space of causal curves from p to q is compact. (A causal curve is a curve, not necessarily smooth, that everywhere proceeds locally within, or on, the future light-cone. Thus, causal curves are timelike curves or curves that are everywhere locally the C^0 *limits* of timelike curves.) Strong causality being assumed, global hyperbolicity is in fact equivalent to the statement that each $I^+(q)\cap I^-(p)$ has compact closure.[7,49] I also remark here that for a causal curve ζ, the set $I^-[\zeta]$ is a TIP if and only if ζ is future-endless,[89] with a similar result holding for TIFs.

Now suppose that $\mathcal{M}$ contains a point q and a future-endless timelike curve γ such that $I^-[\gamma]\subset I^-(q)$. It follows that $\mathcal{M}$ cannot be globally hyperbolic. For if p is a fixed point on γ, and $r_1, r_2, r_3, \ldots$ is a sequence of points proceeding indefinitely far up γ, we obtain a sequence $\zeta_1, \zeta_2, \zeta_3, \ldots$ of causal curves from p to q where ζ_i consists of that segment of γ from p to r_i together with a timelike curve from r_i to q (which exists because $r_i \in I^-[\gamma]$ and $I^-[\gamma]\subset I^-(q)$). If $\zeta_1, \zeta_2, \ldots$ had a limit causal curve ζ, from p to q, then $I^-[\zeta']=I^-[\gamma]$, where $\zeta'=\zeta\cap I^-[\gamma]$ (as is easily seen), whereas ζ' cannot be future-endless, being continued as ζ to the future endpoint q. This would contradict the last statement of the preceding paragraph, showing that ζ does not exist and that $\mathcal{M}$ is consequently not globally hyperbolic.

The converse argument is slightly more technical. I use the notation and proposition numbers of reference 49. Suppose $\mathcal{M}$ is not globally hyperbolic. Then points p and q exist for which $I^+(p)\cap I^-(q)$ does not have compact closure. Hence (propositions 3.9, 5.20) $p\notin \text{int } D^-(\partial I^-(q))$, whereas $p\in I^-(q)$. Consequently (proposition 5.5*h*), $p'\notin D^-(\partial I^-(q))$, where $p'\in I^-(p)$, so there is a future-endless timelike curve γ from p' which does not meet $\partial I^-(q)$. But $p'\in I^-(q)$, whence $\gamma\subset I^-(q)$ and the TIP $I^-[\gamma]$ lies entirely within $I^-(q)$, as is required to prove.

My proposal[57,59] for a *strong cosmic censorship principle* is, therefore, that a physically reasonable classical spacetime $\mathcal{M}$ ought to have the property which can be stated in any one of the following equivalent forms: no TIP lies entirely to the past of any point in $\mathcal{M}$; no TIF lies entirely to the future of any point in $\mathcal{M}$; the TIPs form an achronal set; the

TIFs form an achronal set; $\mathscr{M}$ is globally hyperbolic; there exists a Cauchy hypersurface for $\mathscr{M}$. (This last equivalence is a well-known result due to Geroch.[93]) The plausibility of this rests, of course, on what we expect of a 'physically reasonable' spacetime. Indeed, global hyperbolicity has tended to be regarded by many people as an over-strong restriction. Nevertheless, I believe that one *can* put forward plausibility arguments to support strong cosmic censorship. I next give an indication of this.

To begin with (except in the case of the big bang) it seems not unreasonable to restrict attention, in the first instance, to the case of vacuum only. The reason for this was indicated in section 12.3.1, namely that near 'generic' singularities we expect the Weyl curvature to dominate over the Ricci tensor. This is not totally satisfactory, however, because there are 'cumulative' effects due to the Ricci tensor (namely focussing) which the Weyl tensor can achieve only indirectly through nonlinearities. Nevertheless, the behaviour of vacuum solutions would seem to give a good first approximation near a generic singularity, which avoids the problems raised, for example, by the presence of apparently inessential 'shell-crossing' naked singularities[92] in idealized matter such as 'dust'. As a second approximation one could consider the Einstein–Maxwell theory, for example, which likewise avoids these problems. However, the big bang is a special situation (which relates to its low entropy) and the criterion of 'genericity' need not apply.† But strong cosmic censorship still seems to hold – for different reasons (cf. section 12.3.3). In the singularities of collapse, however, a high-entropy assumption of 'genericity' seems physically reasonable.

The sequence depicted in figures 12.8, 12.9(*a*) and 12.9(*b*) would seem to provide a plausible pattern for the general situation. In figure 12.8 (extended Schwarzschild solution), global hyperbolicity holds, but apparently fortuitously. Each singular TIP intersects a Cauchy hypersurface Σ in a compact region. The data in that region *alone* are all that are needed to imply the existence and nature of the singularity that the TIP represents. But perturb the solution slightly, so that it becomes that of Kerr and extend it maximally in the usual way (as in figure 12.9(*a*)); then the singularity disappears – in the sense that no singular TIP near to the original one now exists, but is replaced by the past $I^-(x)$ of a non-singular interior point x. Thus the original singularity apparently owes its existence to some very special aspect (e.g. the exact spherical symmetry) of the original initial data. Global hyperbolicity is violated in the slightly

† Indeed, it is perfectly in order for the big bang to be what, in the reverse direction in time, would be an *unstable* singularity (cf. section 12.3.3).

perturbed solution, but *because* of this violation there is now a Cauchy horizon $H = H^+(\Sigma)$. The past $I^-(y)$ of a point y of H intersects Σ in a region with *non-compact* closure (extending to infinity). We may take it that the structure of the spacetime (e.g. the curvature) at y is the result of some form of integral of the initial data over this region. If the data are perturbed generically, we are liable to get a *divergence* (owing to the non-compactness and resulting infinite blueshift) so that the non-singular point y changes, in effect, to a singular ideal point with (presumably) diverging curvature.

Suppose, now, that, instead of the asymptotically flat situation we have been examining, we consider a spacelike initial hypersurface Σ that is *compact.* It may still be that, as in figure 12.9(a), certain sets of data on Σ lead to a maximally extended vacuum spacetime which violates global hyperbolicity (e.g. Taub–NUT space $\mathcal{M}$), and a Cauchy horizon $H = H^+(\Sigma)$ is produced. Take $y \in H$, as before, and consider $I^-(y) \cap \Sigma$. This must now have compact closure (since Σ is compact) but it seems that, in a sense, it is liable to be *effectively non-compact* owing to an infinite wrapping round and round Σ by $I^-(y)$ – at least this is what appears to be indicated by an examination of the Bianchi IX models.[80,94] This effective non-compactness would seem to lead to a situation similar to the one just discussed, in which H gets converted to a curvature singularity upon generic perturbation, and strong cosmic censorship holds, because the integrals which define the perturbed curvature on H involve the same data on Σ over and over again, infinitely many times.

One is tempted to conjecture, therefore, that the singularities, like that of figure 12.8, which result from data on an effectively compact region are a 'measure zero' special case, and that while some perturbations give rise to Cauchy horizons, the points on these horizons are liable to be dependent upon effectively non-compact data regions, with infinite blueshift, so that the horizons should be unstable (like that of figure 12.9(a)). In this sense, strong cosmic censorship looks to be very plausibly true† – but the above argument is yet a long way from a proof.

We are thus presented with the picture of a globally hyperbolic universe, which starts from an achronal set $\partial\mathcal{M}$ of initial ideal points (the big bang), then remains topologically unchanging (an implication of global hyperbolicity[93]) – despite the presence of black holes – until the

† From many different viewpoints it is *mathematically* very desirable to be able to restrict attention to globally hyperbolic spacetimes. For example, most of the technical difficulties[89] concerning the topology and identifications for TIP and TIF structure now disappear.

achronal set $\partial\hat{\mathcal{M}}$ of final ideal points is reached. (Strong cosmic censorship implies, in fact, that $\partial\check{\mathcal{M}}$ and $\partial\hat{\mathcal{M}}$ must be regarded as totally disjoint sets.) The initial ideal points are normally taken to be all singular points, but the final ideal points may be either points at infinity or singular points. Points at infinity would normally be thought to arise only in the case of an ever-expanding universe-model, in which case one would also expect singular final ideal points in black holes. But it is also conceivable (though in my view rather unlikely, for reasons similar to those just outlined above) that a universe which recollapses as a whole may nevertheless have certain limited portions that 'escape' to infinity in a (non-singular) ∞-TIP.

In figure 12.12, an indefinitely expanding universe-model is illustrated, showing how $\mathcal{M}$ can remain topologically unchanging – in the sense that $\mathcal{M} \cong \mathbb{R} \times \Sigma$, with each copy of $\mathbb{R}$ a timelike curve and each copy of Σ a spacelike Cauchy hypersurface[93] – even though there may be several

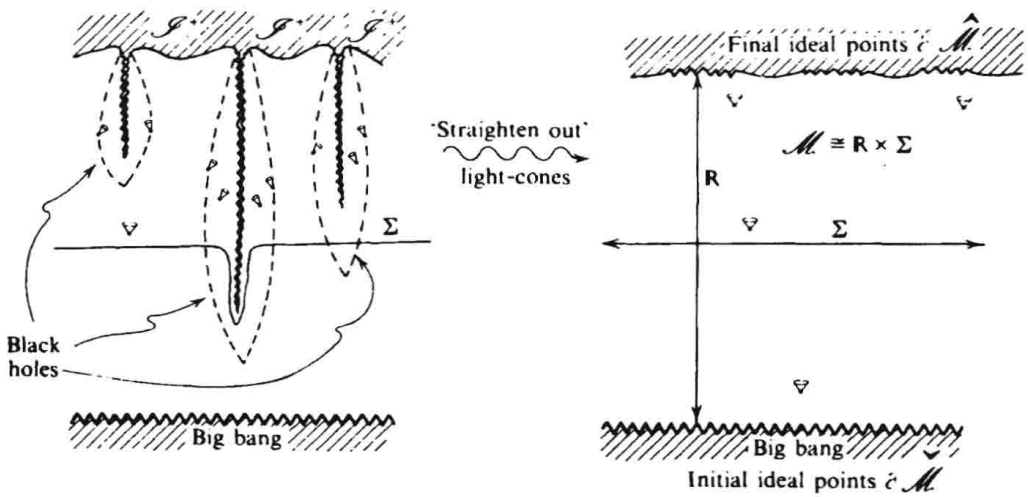


Figure 12.12. A universe-model subject to strong cosmic censorship. The indefinitely expanding case is illustrated. (The different $\mathscr{I}^+$ regions are actually all connected.)

black holes present.[57] The situation for a recollapsing universe-model is similar. The sets $\partial\check{\mathcal{M}}$ and $\partial\hat{\mathcal{M}}$ may or may not have the same topology as Σ, however. (For example, in the Einstein static universe, each of $\partial\check{\mathcal{M}}$ and $\partial\hat{\mathcal{M}}$ is a point, whereas Σ is an S^3.) According to the viewpoint in section 12.3.3, the big bang $\partial\check{\mathcal{M}}$ should, indeed, have the topology of Σ. But it is not at all clear that this should be so of $\partial\hat{\mathcal{M}}$. Another way of phrasing Misner's original hope that the generic Bianchi type IX 'mixmaster' empty models should be free of particle horizons[94] is that $\partial\check{\mathcal{M}}$ should be a

single point.† In time-reversed form, one might have correspondingly expected $\partial\hat{\mathcal{M}}$ to be a single point. But later analysis[95–98] has shown this kind of behaviour to be a very unlikely possibility. Nevertheless, one might anticipate that under certain circumstances $\partial\hat{\mathcal{M}}$ has, in an appropriate sense, fewer than three dimensions (as is the case for the two-dimensional $\partial\hat{\mathcal{M}}$ of a pancake singularity and for Taub space[7]).

If the big bang $\partial\check{\mathcal{M}}$ is, indeed, a smooth spacelike hypersurface with the topology of Σ (cf. section 12.3.3) then it constitutes a very satisfactory initial Cauchy hypersurface. But irrespective of the smoothness of three-dimensionality of $\partial\check{\mathcal{M}}$ and $\partial\hat{\mathcal{M}}$, each can be regarded as providing a home for the ultimate Cauchy data for $\mathcal{M}$ – the ultimate initial data in the case of $\partial\check{\mathcal{M}}$, and the ultimate final data in the case of $\partial\hat{\mathcal{M}}$. Each is, indeed, all-embracing in the sense of being intersected by every endless causal curve in $\mathcal{M}$, and is also achronal. Of course, the exact form that such Cauchy data would take is not yet clear in the general case, but the potentiality is certainly there in principle.

As a final consideration of this section, let us examine the possibility of an initial creation or the final destruction of a charged particle (cf. also section 12.2.7). If the initial creation, at $\partial\check{\mathcal{M}}$, can be regarded as the result of a more-or-less understood process[62,63,66] whereby the curvature itself creates particles, then we would expect not just a single charged particle, but a *pair* of oppositely charged particles. However, it might be that some totally unknown process is involved in particle production at the big bang, whereby charged particles could be produced singly.[67] In this situation, the particle's Coulomb field must also be conjured up at $\partial\check{\mathcal{M}}$, leading, as was remarked in section 12.2.4, to the existence of some effectively source-free incoming radiation.[45] Indeed, such a picture (at first sight, at least) would be presented as the time-reverse of particle destruction at $\partial\hat{\mathcal{M}}$. Imagine a single charge being swallowed by a large spherically symmetrical uncharged black hole. According to the geometry depicted in figure 12.8, this particle must be annihilated alone, on a singular part of $\partial\hat{\mathcal{M}}$. Is it conceivable that this one particle should so disturb the geometry of $\partial\hat{\mathcal{M}}$ that its destruction is held off until finally a particle of the opposite charge is swallowed by the hole and guided in to be annihilated by the first charge, before the energy of both can be absorbed by the singularity? It seems hardly credible. Yet the structure of generic singularities appears to be so delicate and elaborate that even this should not be dismissed out of hand. But the issue that I am attempting to raise here is whether or not

† This is also, and more explicitly, what is demanded for a Sachs–Budic 'deterministic' spacetime.[99]

 Singularities and time-asymmetry. *General relativity an Einstein centenary survey.*
Eds S. W. Hawking and W. Israel, pp. 581–638, Cambridge University Press, 1979

the laws that dominate physical behaviour at a spacetime singularity can, after all, be time-symmetric. It seems to me that those who believe that they can be must face additional serious problems of principle!

12.3.3 A hypothesis and a world-view

Almost all the arguments that I have been presenting seem to focus themselves on one issue: what special geometric structure did the big bang possess that distinguishes it from the time-reverse of the generic singularities of collapse – and why? At this point I should mention the viewpoint of *chaotic cosmology*,[94] which has been in vogue for a number of years. According to this view, the big bang was not initially a very uniform singularity, but came to appear so because dissipative effects (e.g. neutrino viscosity, hadron collisions or particle creation[100–102]) served to iron out all its irregularities. One intended effect of this dissipation is to produce the observed 'high' entropy-per-baryon figure of $\sim 10^9$. Another is to produce the presently observed isotropy. And in order not to impose apparently arbitrary restrictions on the big bang, it was originally conjectured that the chaos of the initial singularity was, in some appropriate sense, *maximal*. In this way a supposedly canonical type of initial state was suggested, whose detailed implications could be worked out and compared with observation.

I have to say, however, that I regard the programme of chaotic cosmology – at least in its 'pure' form of maximal initial chaos – to be basically misconceived. For to assert that the initial chaos is maximal is presumably to assert that the initial *entropy* is maximal. And if this were the case, with time-symmetric physical laws, there would be no time's arrow and therefore no dissipation. (It is no good appealing to the expansion of the universe here – for reasons that were amply discussed in section 12.2.6!)

We might, however, entertain some milder version of chaotic cosmology in which the initial geometric chaos was not maximal but was constrained in some suitable way – the constraints being so adjusted that not only could the observed entropy per baryon be correctly obtained from early dissipation, but that density irregularities adequate for galaxy formation might also be obtainable as a result of a non-uniform initial geometry.[103,105] In my opinion, this version of chaotic cosmology is also quite unsatisfactory. The essential misconception seems to be to regard an entropy per baryon of $\sim 10^9$ as a *high* figure in a gravitational context. Consider a closed recollapsing universe containing, say, 10^{80} baryons.

When an $M_\odot$ black hole forms, it achieves, by the Bekenstein–Hawking formula, an entropy per baryon of 10^{18}, while a $10^{10}\,M_\odot$ galaxy with a $10^6\,M_\odot$ black hole core has an entropy per baryon of 10^{20}. When finally the bulk of this galaxy collapses into the hole, the figure is 10^{28}. But as the collapse of the universe proceeds, these black holes unite into bigger and bigger ones, yielding an immense final entropy per baryon of $\sim 10^{40}$. The entropy resides in the irregularities of the geometry of the final singularity. Reversing the time-sense, we now see what a stupendous entropy would really have been available, had the Creator chosen to make use of it by an initial chaotic geometry. The supposedly 'high' figure of $\sim 10^9$ is small fry indeed! If we are to adopt 'mild' chaotic cosmology as an explanation for the figure 10^9, then we must ask why only the absurdly small fraction of at most $\sim 10^{-31}$ of the 'available chaos' was actually used! (Indeed, the figure would be $\sim 10^{-32}$ if the closed-universe value $\sim 10^{-8}$ for the entropy per baryon were used.)

It would seem that, with such an enormous discrepancy, we should not look for a gravitational explanation† for the figure 10^9. A more hopeful place to look might be in particle physics. I shall return to this question in section 12.4. Likewise it would seem that a *purely* gravitational explanation of the irregularities needed for galaxy formation will not be forthcoming. Again (though with considerably less confidence) I appeal to particle physics – in the very early stages of the expansion.

I propose,[57–60] then, that there should be a complete lack of chaos in the initial *geometry*. We need, in any case, some kind of low-entropy constraint on the initial state. But thermal equilibrium apparently held (at least very closely so) for the *matter* (including radiation) in the early stages. So the 'lowness' of the initial entropy was not a result of some special matter distribution, but, instead, of some very special initial spacetime geometry. The indications of sections 12.2.6 and 12.3.1, in particular, are that this restriction on the early geometry should be something like: *the Weyl curvature C_{abcd} vanishes at any initial‡ singularity.*[58–60]

This hypothesis is still a little vague, and is open to a number of different interpretations. We could require, for example, that the Weyl curvature tend to zero as the initial singularity is approached, or that it should do so at some preassigned faster rate, or, perhaps, that it should

† This is borne out by a recent detailed analysis by Barrow and Matzner[104] who come to the same essential conclusion.

‡ That is, with the notation of section 12.3.2, at $\partial\check{\mathcal{M}}$. Note that this includes the final singular TIF of a Hawking black hole explosion – at which the Weyl curvature indeed vanishes!

Singularities and time-asymmetry. *General relativity an Einstein centenary survey.*
Eds S. W. Hawking and W. Israel, pp. 581–638, Cambridge University Press, 1979

only remain bounded (or merely dominated by the Ricci tensor near the singularity – so that the curvature tensor becomes, in the limit, *proportional* to one whose Weyl part vanishes). I have not examined the differences that might be entailed by adopting different versions of this hypothesis. My inclination is to try, first, the simplest of these, namely that $C_{abcd} \to 0$ (in, say, any parallelly propagated frame), as initial singularities are approached. I shall indicate, shortly, what the rough implications of this would be, though complete details have yet to be worked out.

Let me first explain my view of the role of this hypothesis, as it affects the 'selection' that the Creator might make of *one* particular universe out of the apparently infinite choice available, consistent with given physical laws. Imagine some vast manifold $\mathcal{U}_0$ (where I use the word 'manifold' in a rather loose sense) representing the different possible initial data for the universe, compatible with the physical laws. To select one universe, the Creator simply places a 'pin' somewhere in $\mathcal{U}_0$. But a viewpoint of this article is that we should not be biased in choosing initial rather than final data. So, equally well, the selection of the universe could be envisaged as the Creator's pin being placed in the manifold $\mathcal{U}_\infty$, representing all possible sets of final data compatible with the physical laws. Indeed, one could use any intermediate $\mathcal{U}_t$ representing the possible data 'at time t'.† All are equivalent, the equations of motion (which I am crudely assuming, for the purposes of the present discussion, to be of a classical determinate type) effecting canonical isomorphisms between $\mathcal{U}_0$, $\mathcal{U}_\infty$ and each $\mathcal{U}_t$. Thus we may envisage a *single* isomorphic abstract manifold $\mathcal{U}$ which represents any (or every) one of these, being the set of all possible universe-histories compatible with the physical laws.

We would like to be able to envisage that the Creator's pin is simply placed 'at random' in $\mathcal{U}$ (since if the pin is to be constrained in any further precise way, this would constitute another 'physical law'). But the notion of 'at random' requires that some appropriate *measure* be placed on $\mathcal{U}$. Is 'at random' the same concept when applied to initial conditions as when applied to final conditions? To put the question another way: is the phase-space measure that is naturally defined on $\mathcal{U}_0$ the same as that defined on $\mathcal{U}_\infty$ (or indeed on each $\mathcal{U}_t$) under the isomorphisms? Liouville's theorem tells us that it *is*, provided that we are adopting conventional Hamiltonian physics – and I am not proposing to be

† The concept 'time' is being used very loosely here. In the right-hand picture of figure 12.12 (where indefinite expansion need not be assumed) t could be some suitable parameter ranging from 0 to ∞ and measuring the 'height' up the picture.

'unconventional' in this respect here. I am, furthermore, going to ignore any difficulties† that might arise from a possible infinite-dimensionality, or infinite total measure, of the manifold $\mathcal{U}$ – not to mention all the very serious 'gauge' problems (etc.) which would arise in a proper general-relativistic treatment.

How are we to envisage the entropy of the universe according to this picture? A standard procedure is to coarse-grain each $\mathcal{U}_t$ by dividing it up into compartments, where the different members of any one compartment correspond to states macroscopically indistinguishable from one another at time t. If the Creator's pin pierces $\mathcal{U}_t$ at a point belonging to a compartment of phase-space volume V (measured in units where $\hbar = 1$) then the (Boltzmann) entropy at time t is $k \log V$ (cf. section 12.2.3). This entropy can fluctuate‡ or progressively change with time, because the coarse-grainings of the different $\mathcal{U}_t$ manifolds do not map to one another under the isomorphisms. Low entropy at one time (point in a small compartment) can correspond to high entropy at another (point in a large compartment), where the 'specialness' of the state has now become that of macroscopically indiscernible correlations.

In the present context, however, this description of entropy is not yet satisfactory. The points of $\mathcal{U}$ correspond only to those universe-histories that are compatible with the physical laws at *all* times. It may not be macroscopically discernible at a particular time t whether the laws are satisfied at all other times. I am hypothesizing, here, that there are in fact (local) physical laws which only become important near spacetime singularities, these being asymmetric in time and such as to force the Weyl curvature to vanish at any initial singular point (i.e. point of $\partial\mathcal{M}$). The effect of these laws is that each manifold $\mathcal{U}_t$ turns out to be much smaller than might otherwise have been expected. Only those motions and configurations at time t which are also compatible with the constraints ($C_{abcd} = 0$) at $t = 0$ are allowed. But since these implied constraints on each $\mathcal{U}_t$ are not macroscopically discernible, it is not reasonable, when calculating the entropy at time t, simply to use the phase-space volumes within $\mathcal{U}_t$. Instead, one must consider extended volumes within a certain

† I must apologize, particularly to the experts, for my crude and cavalier treatment of the delicate matters that I am embarking upon. My excuse is that for the questions with which I am now concerned, I do not believe that general-relativistic or thermodynamic sophistication is a key issue.

‡ If we prefer the *ensemble* picture of the world, we can to some extent avoid these fluctuations, envisaging that the Creator uses a *blunt* pin! So long as the diameter of the pin-point is large compared with the coarse-graining, fluctuations are smoothed over.

Singularities and time-asymmetry. *General relativity an Einstein centenary survey.*
Eds S. W. Hawking and W. Israel, pp. 581–638, Cambridge University Press, 1979

larger manifold $\mathcal{W}_t$ – defined in the same way as $\mathcal{U}_t$ but for which these (initial) constraints are not required to hold.

The equations of motion again give isomorphisms between the different $\mathcal{W}_t$ manifolds at different times t (locally at least – and so long as the extra constraining laws remain physically insignificant) and there is a corresponding abstract manifold $\mathcal{W}$ representing the totality of unconstrained universe-histories. The imbedding of $\mathcal{U}$ in $\mathcal{W}$ has a very special relation to the coarse-graining at $t = 0$, because the vanishing of Weyl curvature *is* a macroscopically discernible property. Thus, no $t = 0$ compartment of $\mathcal{U}$ extends outside $\mathcal{U}$, into $\mathcal{W}$. But as t increases, the corresponding coarse-graining compartments of $\mathcal{U}$ extend more and more into $\mathcal{W}$, and accordingly acquire larger and larger extended volumes.† In this way, we regard the 'specialness' of the actual state of the universe (arising from its having started out with $C_{abcd} = 0$) as being more and more of the 'precise-correlations' kind, and less and less of the 'low-entropy' kind, as time progresses.

This is what gives us compatibility with the entropy-increase phenomenon of our universe. In this view, we do not impose any statistical low-entropy assumption at the big bang, but, instead, a precise *local condition* ($C_{abcd} = 0$). Aside from this constraint, there is to be complete randomness, that is, the Creator's pin is placed at random in the manifold $\mathcal{U}$. With this randomness assumption, we can attribute the 'reason' for the absence of initial correlations between particle motions in the initial state (i.e. for the *law of conditional independence*[40]) to the fact that at no time *other* than $t = 0$ is a new local constraint imposed (e.g. there is none imposed at the final singularity $t = \infty$). Correspondingly the 'reason' for the presence of increasing correlations as t increases, and for the increasing entropy, is the initial $C_{abcd} = 0$ constraint. In this way, the problem of time's arrow can be taken out of the realm of statistical physics and returned to that of determining what are the precise (local?) physical laws. I shall briefly discuss this question in section 12.4.

I should make mention, at this point, of the much discussed *anthropic principle*,[106] which is often invoked in connection with the matters I have been raising. This principle would, in effect, imply that the Creator's pin is placed in $\mathcal{U}$, not just at random, but with a weighting factor, weighted in

† Curiously, if it were not for extensions into $\mathcal{W}$, the volumes of the largest compartments would *decrease* with time owing to the profusion of differing macroscopic geometries that would be produced. This seems to correspond to the fact that the universe can get more 'interesting' even though the entropy increases!

favour of universes containing (many?) *conscious observers.* (Furthermore, the pin could also pierce other manifolds $\mathcal{V}$, $\mathcal{W}$, . . . corresponding to all the possible consistent alternative sets of pure-number physical constants, and to all the possible alternative laws of physics. I shall leave aside a discussion of this extended question as being 'beyond the scope of this article'!) Such an anthropic viewpoint has occasionally been invoked in an attempt to explain the entropy imbalance of the observed universe in terms of weighting in favour of a huge 'fluctuation' which might have been needed to produce the conditions necessary for life.[107] The trouble with this is that it is vastly 'cheaper' (in terms of negative entropy) simply to produce a few conscious beings out of some carefully organized particle collisions than it is to build, out of a fluctuation, such entropy imbalance as is familiar on Earth throughout the *entire*, apparently unending, universe – as revealed by the most powerful telescopes!

This is not to say that I believe 'randomness' for the Creator's pin will always remain the best explanation for the state of the world. But, with the addition of the initial $C_{abcd} = 0$ assumption, it seems to work remarkably well, and it saves our having to worry about what 'consciousness' means in physical terms – at least for the time being!

So what, indeed, are the implications of the world-view that I am proposing? With the Weyl curvature initially zero and thermal equilibrium for the matter (and radiation), we shall have something very close to spatial-isotropy and homogeneity for the initial state. Thus the discussions of Friedmann, Robertson and Walker hold good (initially), leading to a striking consistency with various remarkable observations: the uniformity of the 2.7 K black-body radiation (to one part in $\sim 10^3$),[108] the lack of measurable rotation of the universe relative to inertial frames ($< 10^{-16}\,\mathrm{s}^{-1}$),[77] the large-scale gross uniformity of galactic clusters. The big bang singularity itself must have been closely of the Robertson–Walker type.† *Some* fluctuations in the matter distribution are allowed in this view – and, indeed, *must* occur – because the initial restrictions on R_{ab} are statistical, unlike those on C_{abcd}. However, the initial vanishing of C_{abcd} imposes considerable constraints on such initial density and velocity fluctuations. Yet, particle physics will have to be better understood before these fluctuations can be calculated in detail.[109]

White holes are excluded at all times, because their singularities are 'initial' singularities (i.e. points of $\partial\mathcal{M}$) which do not remotely satisfy the constraint $C_{abcd} = 0$. Black holes are allowed, of course, provided that

† This is independently supported by the accuracy of the helium-production calculations.[77,78]

they are formed in the normal way as a consequence of gravitational collapse of a massive body or bodies. But mini-holes are presumably *not* to be expected because they require an initial state with a chaotic geometry. The non-existence of such primordial black holes is consistent with present observations.[110]

Finally, the extraordinary observed behaviour of the entropy of our world – permeating so much of our everyday experience that we tend to take it quite for granted – is the major and most striking consequence of this world-view.

12.4 Asymmetric physics?

Some readers might feel let down by this. Rather than finding some subtle way that a universe based on time-symmetric laws might nevertheless exhibit gross time-asymmetry, I have merely asserted that certain of the laws are not in fact time-symmetric – and worse than this, that these asymmetric laws are yet unknown! But this is not so negative as it might seem. In particular, it tells us to look out for such asymmetries in other places in physics. Where do we look? Ultimately there must be some connection with gravity, since it is the Weyl tensor that describes gravitational degrees of freedom. Classical general relativity is a time-symmetric theory, but one may ask whether this time-symmetry will persist when finally the link with quantum mechanics is appropriately forged. Indeed, if one believes that virtual black holes at the Planck length (10^{-33} cm) are physically important,[73] then the arguments of section 12.2.7 suggest that the vacuum could be time-asymmetric in a significant quantum-gravitational way.

This is not greatly helpful, however, and there is another possible tentative connection between quantum mechanics and gravity that might be more relevant, namely the question of *quantum-mechanical observations*, that was left hanging in section 12.2.2. Certain associations with the Bekenstein–Hawking discussion of entropy should be pointed out. One regards a quantum-mechanical observation to have been 'made', after all, only when something 'irreversible' has taken place. But 'irreversible' here refers to the fact that an essential increase in the entropy has occurred. And entropy, as we have seen, seems to depend on the rather subjective concept of coarse-graining. For a quantum-mechanical observation 'actually' to take place, effecting a real change in the state of the world, one would seem to require *objectivity* for this entropy increase. Now recall that in the Bekenstein–Hawking formula, an entropy measure

is directly put equal to a precise feature of spacetime geometry, namely the surface area of a black hole. Is it that this geometry is now subjective, with the implication that *all* spacetime geometry (and therefore all physics) must in some measure be subjective? Or has the entropy, for a black hole, become objective? If the latter, then may not entropy also become objective in less extreme gravitational situations (cf. section 12.3.1)? Moreover, if it is only with (quantum) gravity that such a passage from subjectivity to objectivity of entropy can occur, then it is with (quantum) gravity that the linear superposition of von Neumann's chain[111] finally fails!

Perhaps, then, the new laws that seem to be needed to extend quantum mechanics (cf. section 12.2.2), so that observations can be incorporated within the theory, constitute some form of *quantum gravity* – by which I mean a theory having quantum mechanics and general relativity as two appropriate limits. I would contend, in any case, that the arguments I have been presenting (notably those of sections 12.2.7 and 12.3.3 which most directly relate to the Bekenstein–Hawking formula) point towards some new theory which is *time-asymmetric.* Accordingly, whatever nonlinear physics† eventually replaces suddenly collapsing wave functions may well turn out to involve an essential time-asymmetry.

At the present state of understanding, such considerations are highly speculative. Yet we know that there *is* a physical law which is time-asymmetric! Somewhere, hidden among the more familiar time-symmetric forces of Nature is one (or perhaps more than one) whose tiny effect has been almost completely masked by these others and has lain undetected in all physical processes bar one: that delicately poised decay of the K^0-meson. I am not suggesting that quantum gravity need be involved here. But, evidently, it is *not* one of Nature's inviolable rules that time-symmetry must always hold!

Moreover, the relative strength of the T-violating (or CP-violating) component in K^0-decay, perhaps 10^{-9}, is possibly suggestive. According to section 12.3.3, we need some explanation from particle physics for the observed entropy-per-baryon number of 10^9. It has, accordingly, been put forward[109,67,62] that perhaps in the early expansion there was a

† An interesting recent suggestion for a nonlinear modification of Schrödinger's equation is that of Bialynicki-Birula and Mycielski[112] according to which an additional term $(b \log |\psi|^2)\psi$ is incorporated. Though not time-asymmetric, there is the link with the present discussion that the constant b (a temperature) has the value $\sim 10^{-8}$ K, which is the Bekenstein–Hawking temperature for a 10 $M_\odot$ black hole. These are the smallest black holes that one has physical reason to believe in.

Singularities and time-asymmetry. *General relativity an Einstein centenary survey.* Eds S. W. Hawking and W. Israel, pp. 581–638, Cambridge University Press, 1979

process of production of baryons and antibaryons, where baryons outnumbered antibaryons by about $1+10^{-9}:1$. Then, in the subsequent annihilation, not only would the (apparently) required nonzero net baryon number be produced, but also the observed entropy per baryon (i.e. photons per baryon). This initial production process, if it were to be explained as arising from something with the vacuum's quantum numbers, would need to violate baryon conservation and (for an imbalance to occur) both *CP*- and *C*-symmetries.† Gross violation of *C* has long been known to be a feature of weak interactions,[20] while *CP*-violation also occurs in the K^0-decay at the apparently required low level. Moreover, baryon non-conservation[76] is *somewhere* to be expected, on the basis of Hawking's black hole evaporation.[56,67]

But to explain the law that $C_{abcd}=0$ initially, we would need something more, namely a violation of each of *T*, *PT*, *CT* and *CPT*. If, for example, *CPT* were not violated, we could take an allowed collapse to a singularity for which $C_{abcd}\neq 0$ and apply the *CPT*-symmetry to obtain a disallowed initial singularity. The symmetries *PT* and *CT* are maximally violated in weak interactions and *T* marginally violated in K^0-decay, but *CPT*-violations have not yet been detected. Of course there is the *CPT*-theorem[113] that lends some theoretical support to universal *CPT*-conservation. But it must be borne in mind that Poincaré covariance is an important assumption of that theorem, whereas I am envisaging situations (singularities, quantum gravity) for which this assumption is explicitly violated. I would contend that somehow, in experimental physics, a *CPT*-violating effect ought eventually to be discernable. But these are early days yet, and it is in no way surprising that such effects have not yet been seen.

I have presented, here, little in the way of quantitative detail. However, many of the phenomena I have been concerned with are of so gross and blatant a nature that much can yet be gleaned even without such detail. The most blatant phenomenon of all these is the statistical asymmetry of the universe. It is, to me, inconceivable that this asymmetry can be present without tangible cause. An explanation by way of the anthropic principle seems very wide of the mark (cf. section 12.3.3). So does an explanation of the 'symmetry-breaking'[114] type, according to which the most probable states of the universe might not share the symmetries of the laws that govern it. (It is difficult to see how our vast universe could

† Or, conceivably, *CPT*-violation in place of *CP*-violation, or *CT* in place of *C*, since there is a time-asymmetry in the universe expansion.

just 'flop' into one or the other of these states when it doesn't even know which temporal direction to start in!) In my own judgment, there remains the *one* ('obvious') explanation that the precise physical laws are actually *not* all time-symmetric!

The puzzle then becomes: why does Nature choose to hide this time-asymmetry so effectively? As we do not yet know the principles that govern Nature's choices of physical law, we cannot yet answer this question. But perhaps we should not be so surprised at a situation in which a fundamental asymmetry lies hidden deep beneath a facade of apparent symmetry. The fauna of this Earth, after all, exhibits with but few exceptions a superficial external bilateral symmetry. How could one have guessed that in the nucleus of every reproducing cell lies a helix whose structure governs the growth of these magnificent symmetrical creatures – yet every one of which is right-handed?

Acknowledgements

I am grateful to Dennis Sciama and, particularly, to Amelia Rechel-Cohn for critically reading the manuscript and drawing my attention to a number of references. My thanks go also to Stephen Hawking for several enlightening conversations.

105
GENERAL INTRODUCTION

Recent Developments in General Relativity, Proc. Roy. Soc. London Ser. A **368** (1979), 3–4

This was the 'general introduction' to a meeting in honour of Einstein's centenary, held at the Royal Society, London.

 General introduction. *Recent Developments in General Relativity, Proc. Roy. Soc. London Ser. A* **368** (1979), 3–4

Proc. R. Soc. Lond. A **368**, 3–4 (1979)
Printed in Great Britain

General introduction

By R. Penrose, F.R.S.
Mathematical Institute, University of Oxford, Oxford, U.K.

It is now over 60 years since Einstein first put forward his General theory of Relativity, providing what was, at that time, a daring new view of the geometry of the world and an extraordinary insight into the nature of the gravitational field – perhaps the single most strikingly original contribution to the scientific thought of recent times. But despite an initial period of active development both in the observational and theoretical sides of the subject, there followed a long period of comparative quiescence during which the subject had seemed to have little contact with the rest of physics and even less with feasible observations or experiments. In recent years, however, this situation has changed dramatically.

On the theoretical side, for example, there has been much clarification of the nature of gravitational waves and of the fact that these waves carry positive mass-energy. Much more is known concerning exact solutions of Einstein's equations. The space–time singularities that had been previously suspected as being features only of special symmetrical solutions are now known to be inevitable for wide classes of space–times. The theory of black holes has arisen and has led to a remarkably complete picture with a detailed and well-understood space–time geometry (the Kerr geometry). The relation to quantum theory, while still enigmatic, has made significant advances (e.g. Hawking radiation). The subject of General-Relativistic astrophysics has been born and has flourished.

Experimental tests of the theory, while originally scanty, are now fairly numerous and seem to be approaching the stage when General Relativity can be considered as a well-tested theory. Indeed, since this discussion meeting was held, new results have been announced (Taylor *et al.* 1979) that indicate that the Binary Pulsar PSR1913+16 is losing energy at just the rate that General Relativity predicts it should, in the form of gravitational radiation. This is perhaps the most impressive test yet of the theory. As regards the building of apparatus for the detection of gravitational waves, remarkable strides have been made, so that what had previously seemed sheer fantasy has (following the initial lead made by J. Weber) become a realistic technological possibility. Finally, with many new observations in astronomy, such as the 3 K background radiation, quasars, radio galaxies, active galactic nuclei, pulsars, X-ray and γ-ray sources, which indicate the presence of strong gravitational fields in the universe, the interaction between theoretical astrophysics and observational astronomy should provide many new testing grounds for General Relativity.

In the papers that follow, most of the important developments in the subject are touched upon, though there are inevitably some omissions (e.g. little was said

about exact solutions or equations of motion at the meeting). We feel that, as we now enter the hundredth anniversary year of Einstein's birth, this meeting provided at least a token appreciation of his most remarkable theory.

Reference (Penrose)

Taylor, J. H., Fowler, L. A. & McCulloch, P. M. 1979 *Nature, Lond.* **277**, 437.

106
GLOBAL GENERAL RELATIVITY

Recent Developments in General Relativity, Proc. Roy. Soc. London Ser. A **368** (1979), 19–21

 Global general relativity. *Recent Developments in General Relativity, Proc. Roy. Soc. London Ser. A* **368** (1979), 19–21

Proc. R. Soc. Lond. A **368**, 19–21 (1979)
Printed in Great Britain

Global General Relativity

By R. Penrose, F.R.S.
Mathematical Institute, University of Oxford, Oxford, U.K.

Much of the theoretical work that has been carried out in General Relativity, particularly in the earlier years of the subject, has been concerned with finding explicit solutions of Einstein's field equations, either in the vacuum case or, with suitable equations of state, when matter is present. These have been very useful in giving us some sort of feeling for the nature of more general 'physically reasonable' solutions, but they can, at best, only be rough approximations to such solutions. Exact solutions must, owing to the limitations of human energy and ingenuity, and to the complexity of Einstein's equations, involve a number of simplifying assumptions, such as special symmetries or particular algebraic forms for the metric or curvature. Sometimes it is legitimate to regard such a special solution as the first term in some perturbation expansion towards something more realistic. But in the highly nonlinear situations of strong gravitational fields, such as in gravitational collapse to a black hole, or perhaps also in cosmology, it is often not clear when the results of such perturbation calculations (themselves often very complicated) can be trusted. High-speed computers can come to our aid (Smarr 1979, this symposium), of course, and can often give important insights in particular situations. But complementary to these are the global qualitative mathematical techniques that have been introduced into relativity over the past several years (Hawking & Ellis 1973; Penrose 1972).

These techniques have, for example, provided rigorous demonstrations (Hawking & Penrose 1970) that Einstein's equations, together with suitable inequalities on the energy–momentum tensor (such as non-negative definiteness of the energy density) can lead inevitably to space–time singularities arising, provided that some qualitative geometric criterion is satisfied. Such a criterion is the existence of some point in the space–time, either whose future light cone or whose past light cone begins to reconverge if one proceeds far enough along the rays. Simple order-of-magnitude considerations lead us to believe that, in suitable situations of gravitational collapse, this criterion will indeed be satisfied (a consequence of Einstein's light-bending effect). Thus, one must infer that, quite independently of any symmetry assumptions, Einstein's equations lead us inevitably to singularities when such a gravitational collapse takes place. Likewise, in a cosmological setting, the criterion can be applied in the reverse direction in time. It is, in principle, an observable effect that the past light cone of the earth, at the present epoch, should begin to reconverge as the big bang is approached. (Objects of given size should, if moved farther and farther away, begin to subtend a *larger* angle again beyond a certain point.) So the singularity of the big bang is, like that of a black hole, a

stable feature of solutions of Einstein's equations. The introduction of irregularities will not make the big bang disappear.

In the case of collapse to a black hole, however, a major unsolved theoretical problem remains with us. While the aforementioned 'singularity theorems' tell us that Einstein's equations predict singularities once a suitable criterion of collapse has been met, we are not yet certain that in physically reasonable situations these singularities must necessarily be hidden from the view of a distant observer, i.e. that they necessarily lie behind the *event horizons* of black holes. This is the still unsolved problem of classical General Relativity referred to as *cosmic censorship* (Penrose 1969, 1973), namely the banishment from our Universe of such naked singularities. In fact, the general qualitative discussion that leads us to believe that black holes *will* result from extreme gravitational collapse depends upon this cosmic censorship hypothesis. If the hypothesis fails, however, then we shall be left with something more alarming even than a black hole, namely a naked singularity.

Though the theoretical evidence that has been accumulating over the years is rather inconclusive, it is my personal opinion that some of this evidence is lending increasing support to some suitable form of cosmic censorship principle actually being a consequence of Einstein's equations. In fact, it seems quite possible that a rather stronger version of cosmic censorship might hold, than had originally been envisaged (Penrose 1974, 1979). This would entail that the singularities of generic collapse could never even be observed by a *local* observer who falls into the black hole – until the observer, himself, directly encounters it. An implication would be that the entire Universe model (assumed classical) must be *globally hyperbolic*, i.e. can be evolved from a single Cauchy hypersurface (Geroch 1970), and that the topology of the Universe (despite the presence of black holes) must be unchanging with time.

An interesting question that arises is: how complete is the analogy between *past* singularities (the big bang, or possibly 'white holes') and *future* singularities (occurring in black holes or in the final singularity of recollapse of a positive-curvature universe)? It seems that a strong case can be presented (Penrose 1979) that some, as yet unknown, physical principle serves to constrain past singularities in a way that does not apply to future singularities. The constraint would be, roughly speaking, that in the immediate neighbourhood of any past singularity the Weyl conformal curvature of the space–time must vanish. Such a constraint would have a number of remarkable consequences, all of which are compatible with present observations. In the first place, the big bang would have to be highly symmetrical, leading to a closely Friedmann–Robertson–Walker form. In the second place, white holes would be definitely excluded; in the third, mini-black holes would not be expected. And, most impressively, there would be a huge imbalance of entropy, leading to the time-asymmetrical behaviour of increasing entropy with which we are all familiar. The physical origin of such time-asymmetrical constraint on space–time singularities is quite obscure, however.

References (Penrose)

Geroch, R. P. 1970 *J. math. Phys.* **11**, 437.

Hawking, S. W. & Ellis, G. F. R. 1973 *The large scale structure of space–time*. Cambridge University Press.

Hawking, S. W. & Penrose, R. 1970 *Proc. R. Soc. Lond.* A **314**, 529.

Penrose, R. 1969 *Riv. Nuovo Cim.*, serie I, vol. 1, numero speciale 252.

Penrose, R. 1972 *Techniques of differential topology in relativity*. Philadelphia: SIAM.

Penrose, R. 1973 *Ann. N.Y. Acad. Sci.* **224**, 125.

Penrose, R. 1974 In *Gravitational radiation and gravitational collapse* (ed. C. DeWitt-Morette). Dordrecht: Reidel.

Penrose, R. 1979 In *General Relativity: an Einstein Centenary survey* (ed. S. W. Hawking & W. Israel). Cambridge University Press.

107
REMARKS ON THE SPARLING AND EGUCHI-HANSON (GOOGLY?) GRAVITONS

Twistor Newsletter **9** (1979). Also published in *Further Advances in Twistor Theory,* Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer. Chapman and Hall/CRC, London, 2001

Remarks on the Sparling and Eguchi-Hanson (Googly?) Gravitons. *Twistor Newsletter* **9** (1979). Also published in *Further Advances in Twistor Theory*, Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer. Chapman and Hall/CRC, London, 2001

§III.4.2 Remarks on the Sparling and Eguchi-Hanson (Googly?) Gravitons

by R. Penrose (TN 9, November 1979)

In §O.3.9, I made the suggestion that the information of the *self-dual* curvature in a solution $\mathscr{M}$ of Einstein's vacuum equation might be stored in a *non-Hausdorff* structure for the deformed twistor space $\mathscr{T}$ relating to the way that the line I is imbedded in its immediate neighbourhood in $\mathbb{P}\mathscr{T}$. This 'googly' structure would compliment the more familiar ('nonlinear graviton') type of deformation in the remainder of $\mathbb{P}\mathscr{T}$ (or $\mathscr{T}$) which stores the information of the *anti-self-dual* curvature of $\mathscr{M}$. It was argued that the presence of gravitational radiation in $\mathscr{M}$, leading to a non-trivial supertranslation between i^0 and i^+, would imply that $\mathbb{P}\mathscr{T}$ would indeed be non-Hausdorff near I. $\mathscr{M}$ was assumed to be analytic with a well-behaved $\mathscr{I}^+$ and i^+, the point i^0 being assumed suitably regular as part of the intrinsic structure of $\mathscr{I}^+$. The two parts of the curvature of $\mathscr{M}$ being appropriately holomorphically stored into the structure of $\mathscr{T}$, one would hope that the entire metric of $\mathscr{M}$ might then be reconstructed, assuming that $\mathscr{M}$ is a free wave without sources, the Einstein vacuum equation being somehow automatically a consequence of the construction.

One difficulty confronting such a programme lies in the fact that there are apparently well-behaved self-dual vacuum solutions in which no such supertranslation conflict arises. A solution of this kind had been provided by Sparling & Tod, (1981) see also §III.1.3, for which the metric can be written as:

$$ds^2 = dudv - dxdy + \lambda^2\Delta^{-3}(udx - xdu)^2, \tag{1}$$

with $\Delta = uv - xy$, $\lambda = const.$ (Thanks to Paul Tod.) Its $\mathscr{I}$ can be seen to have a regular vertex if we introduce

$$U = u\Delta^{-1}, \quad V = v\Delta^{-1}, \quad X = x\Delta^{-1}, \quad Y = y\Delta^{-1} \quad \text{and} \quad \Omega = UV - XY = \Delta^{-1}$$

to obtain the conformally rescaled metric

$$d\hat{s}^2 = \Omega^2 ds^2 = dUdV - dXdY + \lambda^2\Omega(UdX - XdU)^2,$$

the vertex i being $U = V = X = Y = 0$ $(= \Omega)$. This solution appears to be a kind of nonlinear version of the elementary state whose $\mathscr{O}(+)$ twistor function has the

form $\left(\begin{smallmatrix}W\\A\end{smallmatrix}\right)^4 \Big/ \begin{smallmatrix}WW\\CD\end{smallmatrix}$ (or $\mathcal{O}(-)$ twistor function $\left(\begin{smallmatrix}P\\1\\Z\end{smallmatrix}\right)^{-1}\left(\begin{smallmatrix}Q\\1\\Z\end{smallmatrix}\right)^{-5}$). In a similar way, the $\mathcal{O}(+)$ twistor function $\left(\begin{smallmatrix}W\\A\end{smallmatrix}\right)^2\left(\begin{smallmatrix}W\\B\end{smallmatrix}\right)^2 \Big/ \begin{smallmatrix}WW\\CD\end{smallmatrix}$ (or $\mathcal{O}(-)$ function $(\begin{smallmatrix}PQ\\1\,1\\Z\,Z\end{smallmatrix})^{-3}$) seems to lead to

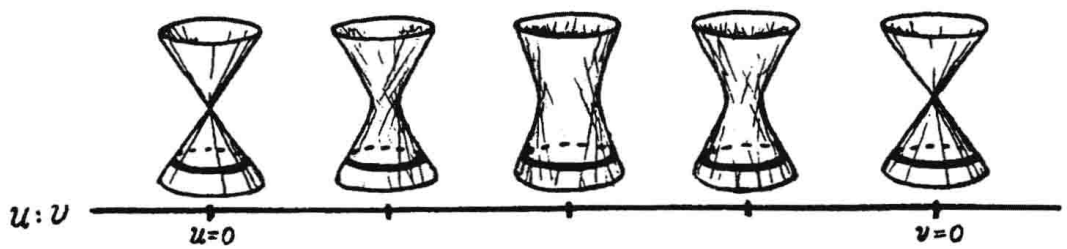
$$ds^2 = dudv - dxdy + \lambda^2\Delta^{-3}(udx - ydv)^2. \tag{2}$$

(Thanks to K.P. Tod.) The same transformation leads to

$$d\hat{s}^2 = \Omega^2 ds^2 = dUdV - dXdY + \lambda^2\Omega(VdY - XdU)^2,$$

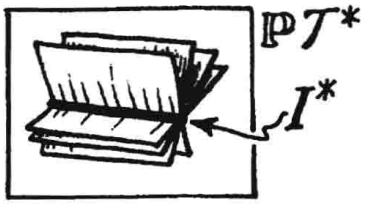
which is just as good as before. G. Burnett-Stuart has shown that the *Eguchi-Hanson* metric can be transformed into the form (2) (see §III.1.3), so the elegant work of (in particular) Hitchin (1979) can be applied in the study of (2). (The E-H instanton has usually been studied in relation to its positive-definite metric, but I shall not be concerned with that aspect here. The existence of a positive-definite metric relates to the fact that in 'Euclideanized' 'pseudo-twistor space', we can arrange that $\begin{smallmatrix}1\\A\end{smallmatrix}$, $\begin{smallmatrix}1\\C\end{smallmatrix}$ and $\begin{smallmatrix}P\\1\end{smallmatrix}$ are complex conjugates, respectively, of $\begin{smallmatrix}1\\B\end{smallmatrix}$, $\begin{smallmatrix}1\\D\end{smallmatrix}$ and $\begin{smallmatrix}Q\\1\end{smallmatrix}$.)

The Hitchin description of the E-H projective (dual) twistor space, which I shall refer to as $\mathbb{P}\mathcal{T}^*$, can be given as follows. Let x, y ,z be coordinates in complex affine 3-space and let $u = \tilde{\pi}_0$, $v = \tilde{\pi}_1$ (no connection with previous u, v ,x ,y). Consider the family of affine quadrics $xy - z^2 = u^2v^2$ for each fixed u,v, where we factor out by the equivalence $(x, y, z, u, v) \equiv (\lambda^2 x, \lambda^2 y, \lambda^2 z, \lambda u, \lambda v)$, so that we get one quadric for each value of $u : v$ (i.e., for each point of dual projective spin-space $\mathbb{S}^*$. These various quadrics provide the standard fibration of $\mathbb{P}\mathcal{T}^*$ over $\mathbb{S}^*$ in the 'nonlinear graviton' construction. In the proper construction for the E-H instanton it is necessary to *blow-up* the vertices of the two *cones* $(xy - z^2 = 0)$ given when $u = 0$ or $v = 0$. However, I shall not be concerned with this aspect of the E-H solution here, but mainly with its behaviour near infinity. The infinity of each Hitchin quadric may be represented as a conic lying on a *projective* version of the quadric. Thus for $\mathbb{P}\mathcal{T}^*$ we have:



Remarks on the Sparling and Eguchi-Hanson (Googly?) Gravitons. *Twistor Newsletter* **9** (1979). Also published in *Further Advances in Twistor Theory*, Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer. Chapman and Hall/CRC, London, 2001

where ◡ represents ∞ on each Hitchin quadric. The union of all these quadrics, but identified with one another at corresponding points of ◡ gives, essentially, $\mathbb{P}\mathcal{T}^*$, where the united ◡ is the line I^*:

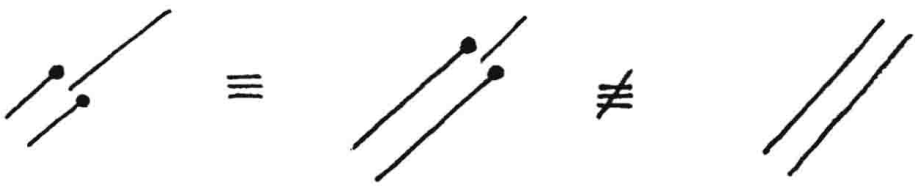


However, as it stands, this is not quite correct. For it is easy to see that ◡ does not have the correct normal bundle in each Hitchin quadric, which should be that of a line in $\mathbb{CP}^2$ — conics in a quadric form a 3-dimensional system and lines in $\mathbb{CP}^2$ a 2-dimensional system. To get the normal bundle right we must take a 2-fold cover of the immediate neighbourhood of ◡ in each Hitchin quadric, the cover being branched along ◡ itself. This takes the square root of the normal bundle of ◡ in each quadric as required. The neighbouring holomorphic curves to ◡ are represented as conics lying on the quadric which *touch* ◡. When the square root is taken, the contact point becomes a simple crossing point for the neighbouring line:

The ambiguity in the square root means that *two* lines arise from each touching conic (near ◡). (Thanks to N.J. Hitchin.)

However, the double cover cannot be extended to the whole (which is simply-connected) in a non-singular way. This is closely related to the fact that the E-H instanton is only ALE (asymptotically *locally* Euclidean) rather than AE. The doubly covered neighbourhood of ◡ can be joined to the remainder of the quadric to provide a *non-Hausdorff* version of the quadric, on the other hand, two points matching to one on the boundary.

When considering such non-Hausdorff features in twistor theory, we should, it would seem, regard as *equivalent* two spaces in which the non-Hausdorff part is simply 'moved along' a little — though not so far as to separate two essential connected pieces completely:

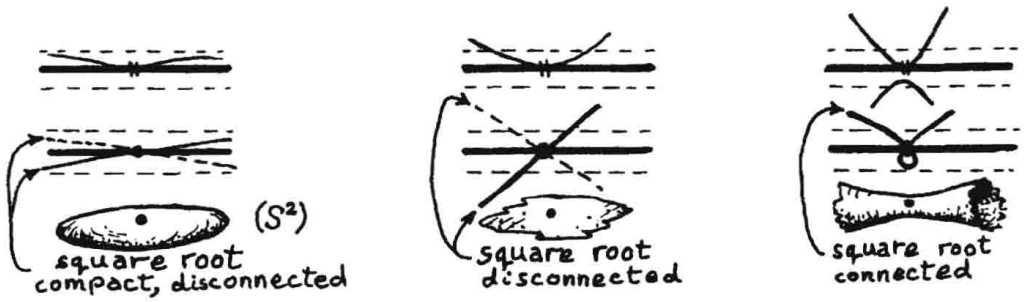


The required concept is, as yet, ill-defined but it relates to the feature already present in twistor theory according to which the location of the 'boundary' of a non-compact complex manifold need not be regarded as significant. Thus we can extend or contract the neighbourhood of $\cup$ in the previous construction and regard the resulting non-Hausdorff manifolds as essentially equivalent.

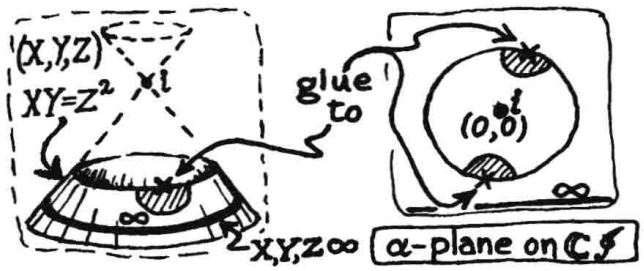
Even if we are concerned only with some small neighbourhood $\mathcal{Q}$ of I^* in $\mathbb{P}\mathcal{T}^*$ (which is itself Hausdorff) some non-Hausdorff features arise when we examine $\mathbb{C}\mathscr{I}$ — as a preliminary to obtaining the 'googly' 'dual' space $\mathbb{P}\mathcal{T}$ of asymptotic twistor lines on $\mathbb{C}\mathscr{I}$. For a point of $\mathbb{C}\mathscr{I}$ is to be interpreted as a 'line' in $\mathcal{Q}$ which meets I^*. *Some* of these 'lines', namely, those close enough to being I^* itself will be compact (within $\mathcal{Q}$) and simply connected S^2's. Others will be non-compact arising as the intersections of $\mathcal{Q}$ with one of the conics on a Hitchin quadric, touching $\cup$. Or, rather, they arise as

Remarks on the Sparling and Eguchi-Hanson (Googly?) Gravitons. *Twistor Newsletter* **9** (1979). Also published in *Further Advances in Twistor Theory*, Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer. Chapman and Hall/CRC, London, 2001

the 2-fold covers of such intersections. But when such a conic get close to being a pair of generators through a point of $\smile$, it turns out that the intersection with $\mathcal{Q}$ is no longer simply connected. Furthermore, this intersection does not have a consistent pair of 'square-roots', the two choices uniting into a single manifold. Thus the space



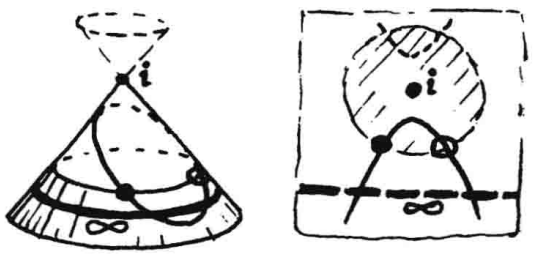
of these 'lines', i.e., $\mathbb{C}\mathscr{I}$, is a non-Hausdorff space. When the square root is ambiguous we get *two* points on $\mathbb{C}\mathscr{I}$ (and this is necessary in order to obtain a 'regular' vertex i). When the square root connects to a single manifold in $\mathcal{Q}$ we get just *one* point on $\mathbb{C}\mathscr{I}$. We can examine this branching behaviour by looking at α-planes on $\mathbb{C}\mathscr{I}$. Each such α-plane corresponds to a Hitchin quadric and is, in a sense, dual to it, or rather to the 2-fold cover of the neighbourhood of $\smile$ corresponding to the intersection with $\mathcal{Q}$. Planes (in the ambient $\mathbb{CP}^3$ of the quadrics) which touch $\smile$ will intersect the quadric in the required conics touching $\smile$. These give points or pairs of points in the required α-plane. Planes touching a conic in $\mathbb{CP}^3$ are represented (by duality) by points of a quadric cone in $\mathbb{CP}^3$. Giving this cone the (affine) equation $XY = Z^2$, where the points at ∞ are to correspond to planes touching the Hitchin quadric at a point of $\smile$ (i.e., which meet the H-quadric in generators), we double cover by choosing $X = \xi^2$, $Y = \eta^2$, $Z = \xi\eta$, providing coordinates $(\xi, \eta) \in \mathbb{C}^2$ for an affine plane. The point $(0, 0)$ corresponds to $\smile$ itself and so gives the point i in the α-plane. If we complete the affine (ξ, η)-plane to a $\mathbb{CP}^2$ we would get the fully 'unwrapped' α-plane which would be flat. But the definition of $\mathbb{C}\mathscr{I}$ arising here does not gives us this $\mathbb{CP}^2$ but a non-Hausdorff version of it. (The 'line at infinity' for the (ξ, η)-plane is singular for the E-H space.) Instead, we obtain a space which consists of a (closed) neighbourhood of i in the (ξ, η)-plane with a boundary which is symmetric in $(0, 0)$, two opposite points of this boundary ((ξ, η) and $(-\xi, -\eta)$) joining to the same place on that portion of the $XY = Z^2$ *cone* that describes the remainder of the α-plane.



The points of $\mathbb{P}\mathcal{T}^*$ (or $\mathcal{Q}$) are described by dual twistor lines on these α-planes that are conics (both on the cone and in the (ξ, η)-plane). The equation of the conic in the (ξ, η) coordinates takes the form

$$A\xi^2 + 2B\xi\eta + C\eta^2 = 1$$

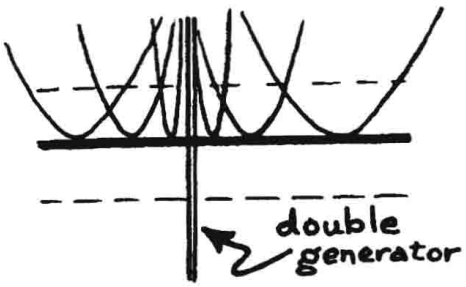
where $\begin{vmatrix} A & B \\ B & C \end{vmatrix} = \text{const.}$ (Thanks to Paul Tod and George Burnett-Stuart.)



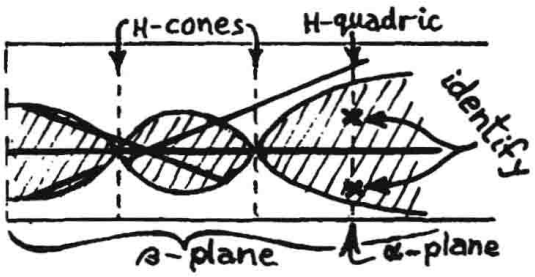
Each choice of the conic on the cone can given rise to *two* alternative extensions into the (ξ, η)-plane, providing the two different points of $\mathcal{Q}$ that correspond to a single point of a Hitchin quadric.

The foregoing description refers only to α-planes for which the Hitchin quadric is not a cone. In the *cone* cases, instead of having a pair of generators through each point of ⌒ we have a single double generator. When conics touching ⌒ approach such a generator, their intersections with $\mathcal{Q}$ need never become multiply connected and so they can have a *disconnected* square root throughout.

 Remarks on the Sparling and Eguchi-Hanson (Googly?) Gravitons. *Twistor Newsletter* **9** (1979). Also published in *Further Advances in Twistor Theory*, Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer. Chapman and Hall/CRC, London, 2001



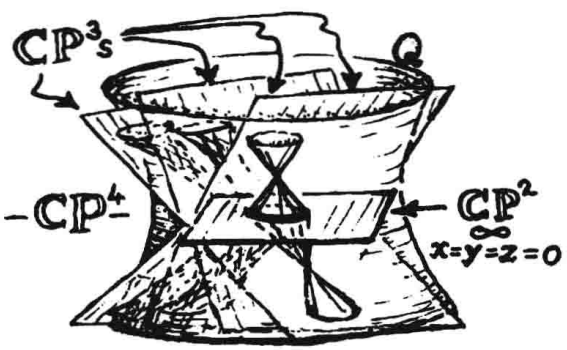
Thus the corresponding α-plane is simply a standard Hausdorff $\mathbb{CP}^2$. (It is not altogether clear whether or not an additional non-Hausdorffness should be introduced, making the ∞-conic on the $XY = Z^2$ cone double on the other α-planes, corresponding to the fact that generator pairs are reducible and consequently the double generators might also count as pairs of points on their associated α-planes. I shall ignore this possibility here.) We now have, for the structure of $\mathbb{CS}$, a non-Hausdorff space which is an identified version of the standard flat-space $\mathbb{CS}$, all but two of the α-planes on $\mathbb{CS}$ exhibiting this non-Hausdorffness. To find the *twistor* lines, lying in β-planes on $\mathbb{CS}$, we take *straight* lines in these β-planes, but when these lines enter the identified regions, they must be represented as *pairs* of lines, corresponding under the identification map $((\xi, \eta) \to (-\xi, -\eta))$.



Each time such a line pair enters an unidentified region, there is a *choice* as to which of the two possible continuations to take up. Thus, the space $\mathbb{PT}$ of all such twistor lines branches and will be a non-Hausdorff space, the branching occurring when the number of disconnected intersections of the straight line with the unidentified region increases.

The upshot of all this is that $\mathbb{P}\mathcal{T}$ is indeed a non-Hausdorff space whose structure contains the essential information required (namely, the location of two planes — corresponding in the linear limit to $\frac{P}{Z} = 0$, $\frac{Q}{Z} = 0$). This ought to be encouraging, but as far as I have been able to ascertain, the non-Hausdorff region does not reach the immediate neighbourhood of I in $\mathbb{P}\mathcal{T}$ as would seem to be required, although it can be extended to do so. There may well be further subtleties to be uncovered. The status of all this is problematical!

The Sparling solution is similar to that of E-H except that the two Hitchin cones coincide. The following representation is very helpful in the study of both solutions. Choose Hitchin coordinates x,y,z,u,v as before, but put $w = u^2$ and $t = v^2$. Then we have projective coordinates (x, y, z, w, t) for $\mathbb{CP}^4$ and a 3-quadric $Q : xy - z^2 - wt = 0$ lying in it. The special 2-plane $x = y = z = 0$ defines $\cdots$ as its intersection with Q. The 3-planes through this 2-plane cut Q in a family of quadrics which are the Hitchin quadrics, each counted twice over (with the exception of the two cones which arise when the 3-planes touch Q — they count once each). The ambiguity arises from $(u, v) \mapsto (u, -v)$.



The Sparling solution is similar except that Q is now a cone with equation $xy - z^2 - w^2 = 0$. The *one* Hitchin cone joins $\cdots$ to the vertex $(0, 0, 0, 0, 1)$, the remaining Hitchin quadrics being non-singular. For both solutions, $\mathbb{P}\mathcal{T}$ can be interpreted, apart from multiple coverings and non-Hausdorffness, as a certain family of 3-quadrics with line vertices, these lines being tangent lines to $\cdots$. The *space* $\mathbb{P}\mathcal{T}^*$, apart from multiple coverings, etc. is, of course, the quadric Q.

458 Remarks on the Sparling and Eguchi-Hanson (Googly?) Gravitons. *Twistor Newsletter* **9** (1979). Also published in *Further Advances in Twistor Theory*, Vol. 3. Eds L. J. Mason, L. Hughston, P. Z. Kobak, and K. Pulverer. Chapman and Hall/CRC, London, 2001

References

Hitchin, N.J. (1979) Polygons and Gravitons, Proc. Camb. Phil. Soc., **85**, 465–476.

Sparling, G.A.J. & Tod, K.P. (1981) An example of an $\mathscr{H}$-space, J. Math. Phys., **22**, no. 2, 331-2.

108
THE ANTI-SELF-DUAL COULOMB FIELD'S NON-HAUSDORFF TWISTOR SPACE

Twistor Newsletter **9** (1979). With G. A. J. Sparling. Also published in *Further Advances in Twistor Theory,* Vol. 1. Eds L. J. Mason and L. P. Hughston, Longman, 1990

This non-Hausdorff feature of the twistor space of the Coulomb field underlies the general twistor treatment of charged world-line sources, as developed later by Toby Bailey; see Chapters 134 and 135 and *Further Advances in Twistor Theory,* Volume 1 (eds. L.J. Mason and L.P. Hughston), Longman, 1990. If we want the patchwork of Chapter 73 to apply globally, we need to take this non-Hausdorffness into account.

 The anti-self-dual Coulomb field's non-hausdorff twistor space. *Twistor Newsletter* **9** (1979). With G. A. J. Sparling. Also published in *Further Advances in Twistor Theory*, Vol. 1. Eds L. J. Mason and L. P. Hughston, Longman, 1990

18. The Anti-self-dual Coulomb Field's Non-Hausdorff Twistor Space.

The global twistor description of the Coulomb field (or, rather, of its anti-self-dual part) in $\mathbb{CM}$ (complex compactified Minkowski space) encounters certain difficulties owing to the fact that it is a 2-valued field: the field changes sign as $\mathcal{P}$ is crossed. Furthermore, deleting $\mathcal{P}$ or $\mathbb{CP}$ does not help since, within $\mathbb{CM}$, $\mathbb{CP}$ can be "dodged around". A further difficulty is that the charge world-line $\mathcal{Q}$, which is the only invariant structure available, defines merely a quadric Q in $\mathbb{PT}$, and the difference $\mathbb{PT}-Q$, being Stein, has no holomorphic (coherent) cohomology and no useful holomorphic line bundles over it. Moreover the line bundles over $\mathbb{PT}$ itself are simply powers of the Hopf bundle $\mathcal{H}$. Yet, over any reasonable neighbourhood of a general line in $\mathbb{PT}$ there will be a "twisted photon" line bundle representing the field. It would seem that one ought to be able to piece these line bundles together to obtain a line bundle representing the field globally. What can this global line bundle be?

The answer is that it is a bundle over the non-Hausdorff complex manifold $\mathbb{PT}_Q$ which is obtained by taking two copies of $\mathbb{PT}$ and identifying them along corresponding points of $\mathbb{PT}-Q$. Thus, $\mathbb{PT}_Q$ is like $\mathbb{PT}$ except that each point of Q in $\mathbb{PT}$ becomes a pair of Hausdorffly non-separated points in $\mathbb{PT}_Q$. To construct the Coulomb bundle $\mathcal{C}$ over $\mathbb{PT}_Q$, we patch together the trivial 0-bundle $\mathcal{H}^0$ over one copy of $\mathbb{PT}$ with the (-2)-bundle $\mathcal{H}^2$ over the other copy of $\mathbb{PT}$ as follows. Let $Z \in \mathbb{PT}$ denote the projective twistor corresponding to $\underline{Z} \in \mathbb{T}$. Then the points of the product bundle $\mathcal{H}^0$ are simply pairs (λ, Z), with $\lambda \in \mathbb{C}$, $Z \in \mathbb{PT}$, whereas the points of $\mathcal{H}^2$ can be represented as $(\underline{I}\underline{Z}, Z)$. Let the equation of Q be $\underline{Q}_{\underline{Z}\underline{Z}} = 0$. Then the fibres over points of $\mathbb{PT}-Q$ (i.e. where $\underline{Q}_{\underline{Z}\underline{Z}} \neq 0$) are obtained by simply identifying (λ, Z) with $(\underline{I}\underline{Z}, Z)$ where $\lambda = \underline{Q}_{\underline{Z}\underline{Z}}$. (This is clearly linear along the fibre as is required). Over Q itself we retain two separate fibres over each point, so the result is a line bundle over $\mathbb{PT}_Q$.

To work out the field at a point p of $\mathbb{CM}$, take the corresponding line l in $\mathbb{PT}$. Provided that p is not zero perpendicular distance from $\mathcal{Q}$, l will meet Q in two distinct points. Consider one of these points to be taken from one copy of Q in $\mathbb{PT}_Q$ and the other to be taken from the other copy. The part of the bundle $\mathcal{C}$ which lies above l described in this way is a 1-bundle ($\mathcal{H}'$), and the same holds if we vary l a little bit. This non-trivial structure over a neighbourhood of l gives the (a.s.d.) Coulomb field (for unit charge) in the corresponding neighbourhood of p. The field blows up when the two intersections of l with Q coincide. The field changes sign when the roles of the two intersections are interchanged. This may be achieved continuously, varying l with the points never coinciding.

To construct the charge n (a.s.d.) Coulomb field, identify $\mathcal{H}^0$ with $\mathcal{H}^{2n}$ over $\mathbb{PT}-Q$, where $(\lambda, Z) \equiv (\underbrace{\underline{I}\underline{Z}\cdots\underline{I}\underline{Z}}_{2n}, Z)$ with $\lambda = (\underline{Q}_{\underline{Z}\underline{Z}})^n$. (Tensor through by $\mathcal{H}^{1-n}$ at the end, if desired.) Reference: Advances in Twistor Theory §3.2, §3.3. ~George Sparling and Roger Penrose

109
ON THE TWISTOR DESCRIPTIONS OF MASSLESS FIELDS

Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978). Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

This is an amalgamation of the material in Chapters 77, 78, 82, 88, 89, and 102, with occasional additional illustrations and changes in wording. It provides a useful overview of the very basic ingredients of the input of sheaf-cohomological ideas into twistor theory.

462 On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

R Penrose

On the twistor descriptions of massless fields

1. TWISTOR FUNCTIONS AND SHEAF COHOMOLOGY *

Recall [1,2] the following properties of a twistor function $f(Z^\alpha)$, to be used for generating a zero rest-mass field $\phi_{A'\ldots L'}$, or $\phi_{A\ldots K}$, by means of a contour integral,

$$\phi_{A'\ldots L'} = \frac{1}{(2\pi i)^2}\oint \pi_{A'}\ldots\pi_{L'}\, f(Z^\alpha)d^2\pi,$$

$$\phi_{A\ldots K} = \frac{1}{(2\pi i)^2}\oint \frac{\partial}{\partial\omega^A}\ldots\frac{\partial}{\partial\omega^K} f(Z^\alpha)d^2\pi:$$

(i) The function f is holomorphic on some domain $\mathcal{D}$ of twistor space not invariant under SU(2,2) (nor under the Poincaré group, nor the Lorentz group).

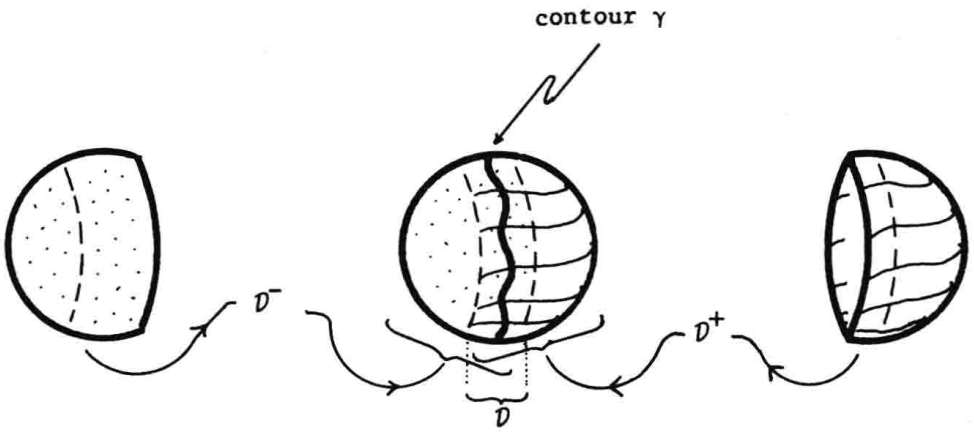
(ii) There is a "gauge" freedom G whereby

$$f \longmapsto f + h^- + h^+,$$

where $h^\pm$ is holomorphic on some extended domain $\mathcal{D}^\pm(\supset \mathcal{D})$ of twistor space in which the contour γ can be deformed to a point (to the "left" in $\mathcal{D}^-$ and to the "right" in $\mathcal{D}^+$)

*Taken (with minor corrections) from Twistor Newsletter 2 (10 June, 1976).

On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

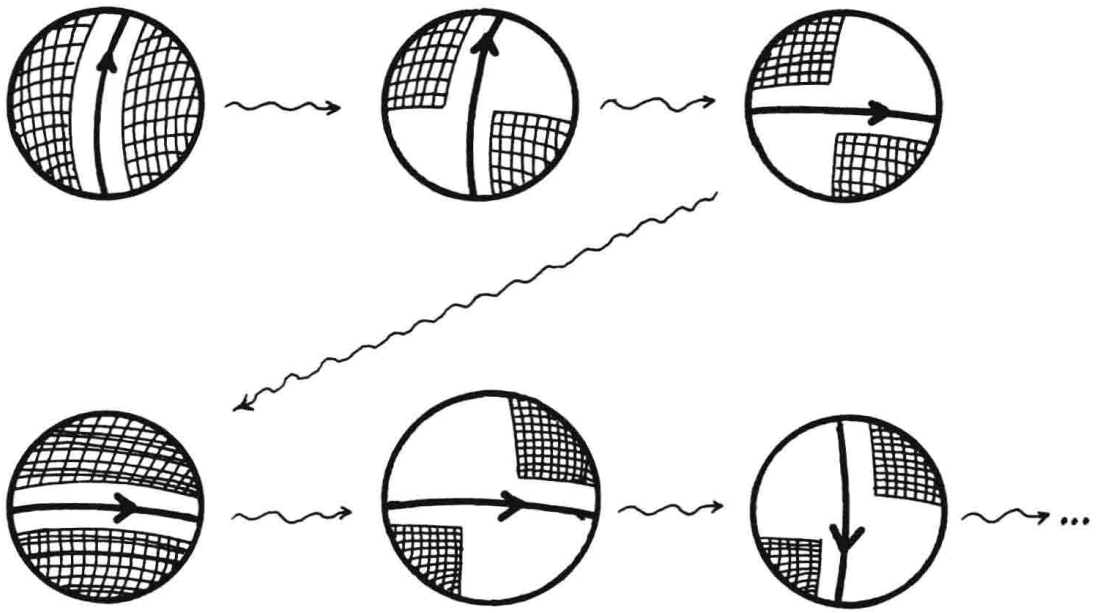


Domain $\mathcal{D}^- \cup \mathcal{D}^+$ (e.g. $\mathbb{T}^+$) symbolically represented as S^2.

(iii) G depends on the location of γ (and γ is not invariant under SU(2,2)--nor under the Poincaré group nor the Lorentz group--so G is not invariant either).

(iv) By invoking G, then moving γ, then invoking a new G, moving γ again,

 On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91



we can obtain a whole family of equivalent twistor functions, all giving the same field, the entire family being invariant under SU(2,2).

(v) This seems a little nebulous, and, for example, how do we add two such families (to give $\phi_{A'\ldots L'} + \psi_{A'\ldots L'}$), etc. etc?

A new viewpoint concerning twistor functions has been gradually emerging which makes good mathematical sense of all this: a twistor function is really to be viewed as a representative function (or cocycle) defining an element of a sheaf cohomology group. Now, the twistor theorist, when attacked by a purist for shoddiness in the domains, can counter-attack armed with his sheaf!

Thumbnail sketch of (relevant) sheaf cohomology theory

First let us recall how ordinary (Čech) cohomology works. Let X be a space (a Hausdorff paracompact topological space, say). Cover X with a locally finite system of open sets U_i. We define a cochain (with respect to this

covering) with coefficients in an additive abelian group $\mathbb{G}$ (say, the integers $\mathbb{Z}$, the reals $\mathbb{R}$, or the complex field $\mathbb{C}$) in terms of a collection of elements f_i, f_{ij}, f_{ijk}, ... $\in \mathbb{G}$, assigned to the various U_i and their non-empty intersections: f_i assigned to U_i; f_{ij} assigned to $U_i \cap U_j$; f_{ijk} assigned to $U_i \cap U_j \cap U_k$..., and $f_{ij} = -f_{ji}$, $f_{ijk} = =f_{jik} = f_{jki} = \ldots, \ldots,$ i.e. $f_{i\ldots\ell} = f_{[i\ldots\ell]}$. Then

$$0\text{-cochain } \alpha = (f_1, f_2, f_3, \ldots)$$

$$1\text{-cochain } \beta = (f_{12}, f_{23}, f_{13}, \ldots)$$

$$2\text{-cochain } \gamma = (f_{123}, f_{124}, \ldots)$$

(where $U_1 \cap U_2$, $U_2 \cap U_3$, $U_1 \cap U_3$, ... , $U_1 \cap U_2 \cap U_3$, $U_1 \cap U_2 \cap U_4$,... are the non-empty intersections of U_i's).

Define <u>coboundary operator</u> δ as follows:

$$\delta\alpha = (\underbrace{f_2 - f_1}_{\text{"}f_{12}\text{"}}, \underbrace{f_3 - f_2}_{\text{"}f_{23}\text{"}}, \underbrace{f_3 - f_1}_{\text{"}f_{13}\text{"}}, \ldots)$$

$$\delta\beta = (\underbrace{f_{12} - f_{13} + f_{23}}_{\text{"}f_{123}\text{"}}, \underbrace{f_{12} - f_{14} + f_{24}}_{\text{"}f_{124}\text{"}}, \ldots)$$

etc.

(where, again, $U_1 \cap U_2$,... are the non-empty intersections). Then we have $\delta^2 = 0$. We call γ a <u>cocycle</u> if $\delta\gamma = 0$; we call γ a <u>coboundary</u> if $\gamma = \delta\beta$ for some β. Define the <u>p^{th} cohomology group</u> by:

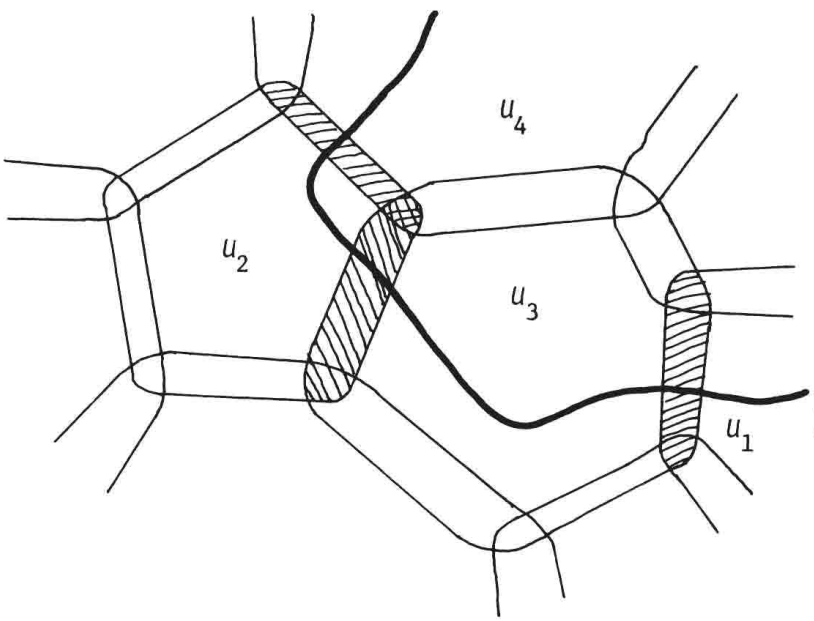
$$H^p_{\{U_i\}}(X, \mathbb{G}) = \left(\begin{matrix}\text{additive group} \\ \text{of p-cocycles}\end{matrix}\right) \Big/ \left(\begin{matrix}\text{additive group} \\ \text{of p-coboundaries}\end{matrix}\right).$$

<u>Note</u>: $H^p_{\{U_i\}}(X, \mathbb{G})$, as defined, depends on the covering $\{U_i\}$. What we should do, to define $H^p(X, \mathbb{G})$, is to take the appropriate "limit" of all these

 On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

$H^p_{\{U_i\}}(X,\mathcal{G})$ for finer and finer coverings $\{U_i\}$ of X. However (for X suitably non-pathological) we can always settle on a particular "sufficiently fine" covering $\{U_i\}$ where, in effect, there is no "relevant topology" left on each U_i or intersections thereof (i.e. all the $H^p(U_i \cap \ldots \cap U_k, \mathcal{G})$ vanish for all $p > 0$--although, as it stands, this is somewhat unhelpful because a direct limit is then already involved in the definition of "sufficiently fine"). Then this $H^p_{\{U_i\}}(X,\mathcal{G}) = H^p(X,\mathcal{G})$. I shall henceforth assume that such a "sufficiently" fine covering has been taken, and that it is countable and locally finite.

Now what does this definition have to do with the familiar "dual" relation to ordinary homology $H_p(X,\mathcal{G})$? How does γ assign values (elements of $\mathcal{G}$) in a linear fashion to p-cycles in X, where γ is some element of $H^p(X,\mathcal{G})$? Intuitively:



Add together the f_{ij} for shaded regions (entered by 1-cycle); similarly for higher dimensional p-cycles. (A complete rigorous discussion can also be given.)

1-cycle κ (closed curve)

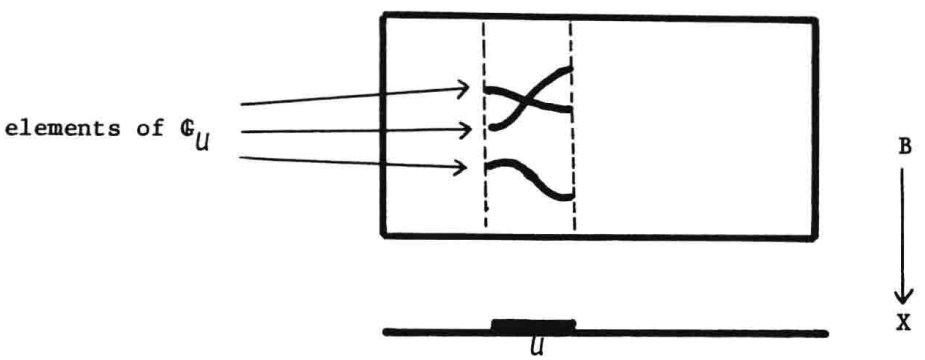
If γ is defined by $(f_{12}, f_{23}, \ldots)$, then $\gamma(\kappa) = f_{42} + f_{23} + f_{31} + \ldots$ (correctly signed) f's on regions entered by κ.

What about sheaf cohomology though? It's really a rather natural generalization of the above. But first we must rephrase the concept of a cochain slightly. Rather than thinking of f_i as simply an element of $\mathbb{G}$ "assigned" to U_i, and f_{ij} "assigned" to U_{ij}, etc., we think of f_i as a function defined on U_i which happens, in the above, to take this constant value "f_i" $\in \mathbb{G}$, and we think of f_{ij} as a constant function on $U_i \cap U_j$ with values in $\mathbb{G}$, etc. (This has the incidental advantage that the requirement $U_i \cap U_j \neq \emptyset$ is now unnecessary.) This is still ordinary (Čech) cohomology. But the generalization to sheaf cohomology is now easily made: the functions f_i, f_{ij}, f_{ijk},... are now not required to be constant. (In fact, we could even allow that the additive group $\mathbb{G}$, in which the values of the f's reside, may vary from point to point in X, or, indeed, the f's need not really be "functions" in the ordinary sense at all--but such situations will not be considered here.) We may require that the f's be restricted in some way--in particular, for the purposes of the present applications we shall often require the f's to be holomorphic (with, here, $\mathbb{G} = \mathbb{C}$) and X a complex manifold--or, we may consider other related classes of functions.

So, what's a sheaf? Actually, I shan't even bother with a formal definition (which can be found in, for example, the references [3,4,5]). The essential point is that a sheaf is so defined that the Čech cohomology works just as well as before. In fact, a sheaf S defines an additive group $\mathbb{G}_U$ for each open set $U \subset X$. For example, $\mathbb{G}_U$ might be the additive group of all holomorphic functions on U (taking X to be a complex manifold). In this case we get the sheaf, denoted $\mathcal{O}$, of germs of holomorphic functions on X. Slightly more generally, we might consider "twisted" holomorphic functions, i.e. functions whose values are not just ordinary complex numbers, but taken in some complex line bundle over X (think of "spin-weighted" functions, for

 On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

example). An important example of such a twisted function would arise if X were taken to be projective twistor space $\mathbb{P}\mathbb{T}$ (or a suitable portion thereof) and the functions considered were to be homogeneous (and holomorphic) of some fixed degree n in the twistor variable. For each open set $U \subset X$ we take $\mathcal{C}_U$ to consist of all such twisted functions on U, and the resulting sheaf denoted $\mathcal{O}(n)$, is called the "sheaf of germs of holomorphic functions twisted by n" (on X). More generally we might consider functions whose values lie in some vector bundle B over X (e.g. we might consider tensor fields on X) and $\mathcal{C}_U$ would consist of the cross-sections of the portion of B lying above U:



Cochains are defined as before (with $f_i \in \mathcal{C}_{U_i}$, $f_{ij} \in \mathcal{C}_{U_i \cap U_j}$, ...) and the coboundary operator δ, just as before. Then we obtain the p^{th} cohomology group of X, with coefficients in the sheaf S, as

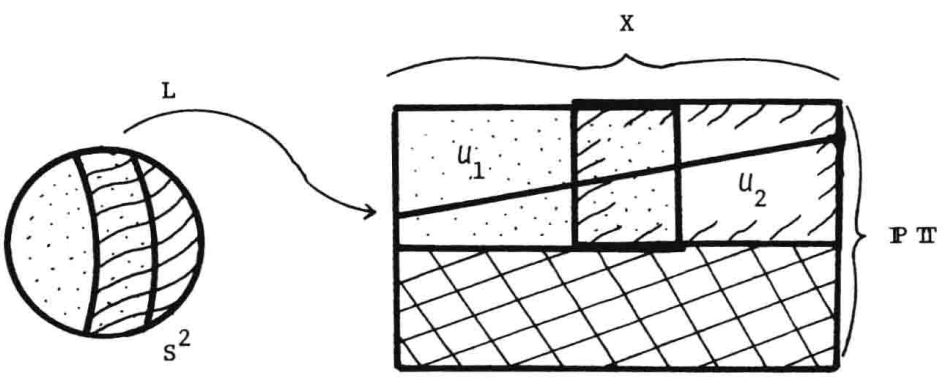
$$H^p(X,S) = \left(\begin{array}{c}\text{p-cochains with}\\ \text{coefficients in } S\end{array}\right) \Big/ \left(\begin{array}{c}\text{p-coboundaries with}\\ \text{coefficients in } S\end{array}\right).$$

On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

As before, we would need to take the appropriate limit for finer and finer coverings $\{U_i\}$ of X, but we can settle on one "sufficiently fine" covering if desired. Provided $\mathcal{S}$ is what is called a coherent analytic sheaf (and we are interested primarily in such sheaves--locally defined by n holomorphic functions factored out, if desired, by a set of s holomorphic relations), then "sufficiently fine" can be taken to mean that each of U_i, $U_i \cap U_j$, $U_i \cap U_j \cap U_k$,... is a Stein manifold [4] (and it is sufficient just to specify that each U_i is Stein). In effect, a Stein manifold is a holomorphically convex open subset of $\mathbb{C}^n$ (or a domain of holomorphy). If X is Stein and $\mathcal{S}$ coherent, then $H^p(X,\mathcal{S}) = 0$, if $p > 0$. Note: $\mathcal{O}$ and $\mathcal{O}(n)$ are coherent.

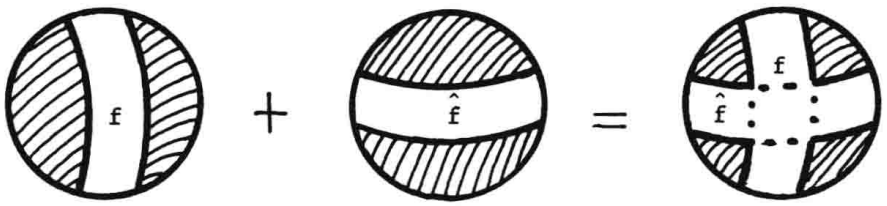
Twistor functions as elements of $H^1(X,\mathcal{O}(n))$

Let X be some suitable portion of projective twistor space $\mathbb{P}\mathbb{T}$, say some neighborhood of a line in $\mathbb{P}\mathbb{T}$ (corresponding to some neighborhood of a point in Minkowski space), or say $\mathbb{P}\mathbb{T}^+$, or $\overline{\mathbb{P}\mathbb{T}^+}$. Suppose we can cover X with two sets U_1, U_2 (each open in X) such that every projective line L in X meets $U_1 \cap U_2$ in an annular region and where $U_1 \cap U_2$ corresponds to the domain of definition of some twistor function $f(Z^\alpha)$, homogeneous of degree n in the twistor Z^α.



Then $f = f_{12}$ is a twisted function on $U_1 \cap U_2$. There are no other $U_i \cap U_j$'s, so f_{12} by itself defines a 1-cochain β, with coefficients in $\mathcal{O}(n)$, for X. Clearly $\delta\beta = 0$, so β is a cocycle. The 1-coboundaries, for this covering, are functions of the form $h_2 - h_1$, where h_2 is holomorphic on U_2 and h_1, on U_1. Calling $\mathcal{U} = U_1 \cap U_2$, $\mathcal{U}^- = U_1$, $\mathcal{U}^+ = U_2$, $h^- = -h_1$, and $h^+ = h_2$, we observe that the "equivalence" between twistor functions under the "gauge" freedom G that we started out with is just the normal cohomological equivalence between 1-cochains β, β' that their difference be a coboundary: $\beta' - \beta = \delta\alpha$, with $\alpha = (h_1, h_2)$. This suggests that we view the twistor function f as really defining us an element of $H^1(X, \mathcal{O}(n))$.

But this is all with respect to a particular covering of X, namely by $\{U_1, U_2\}$. Is this covering "fine" enough? There actually is a technical problem here. We cannot (normally) arrange that U_1 and U_2 are Stein manifolds (and in those exceptional cases when we can so arrange this, we would lose the invariance properties that we are striving for). It turns out, in fact, that this problem is not serious. One can show by direct construction (using the inverse twistor function--in the cases $n \leq -2$ at least, but probably in all cases) that for any <u>given</u> (analytic, positive frequency) field such a covering by two such sets U_1, U_2 is sufficient in the case $X = \overline{\mathbb{P}\mathbb{T}^+}$. Note that though X is invariant under SU(2,2), in this case, the <u>covering</u> is not. However, the cohomology group $H^1(X, \mathcal{O}(n))$ <u>is</u> invariant. Let us illustrate this by adding two elements of $H^1(X, \mathcal{O}(n))$ one of which is defined by a twistor function f, with respect to the covering $\{U_1, U_2\}$, and the other by $\hat{f}$, with respect to $\{\hat{U}_1, \hat{U}_2\}$, the second covering being a rotated version of the first. Schematically:



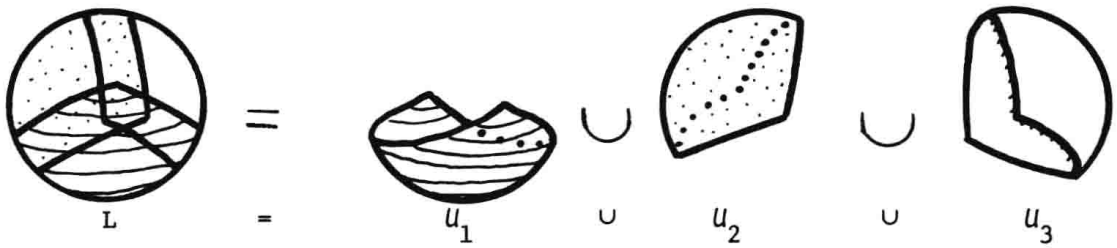
We define a representative cochain for the sum by taking the common refinement of both coverings. (Denote $\hat{\hat{U}}_1 = U_1 \cap \hat{U}_1$, $\hat{\hat{U}}_2 = U_1 \cap \hat{U}_2$, $\hat{\hat{U}}_3 = U_2 \cap \hat{U}_1$, $\hat{\hat{U}}_4 = U_2 \cap \hat{U}_2$ to give the refined covering $\{\hat{\hat{U}}_1, \hat{\hat{U}}_2, \hat{\hat{U}}_3, \hat{\hat{U}}_4\}$.) The 1-cocycle $\beta + \hat{\beta} = \hat{\hat{\beta}}$ is

$$(\hat{\hat{f}}_{12}, \hat{\hat{f}}_{13}, \hat{\hat{f}}_{23}, \hat{\hat{f}}_{14}, \hat{\hat{f}}_{24}, \hat{\hat{f}}_{34}) = (\hat{f}, f, f-\hat{f}, f+\hat{f}, f, \hat{f}).$$

Because of the "direct construction" argument mentioned above, this 1-cocycle will be cohomologous to (i.e. differing by a coboundary from) a cocycle of the form $(0, \hat{\hat{f}}, \hat{\hat{f}}, \hat{\hat{f}}, \hat{\hat{f}}, 0)$, so we can refer it back to the original covering if desired (although with U_1 and U_2 perhaps reduced slightly in size).

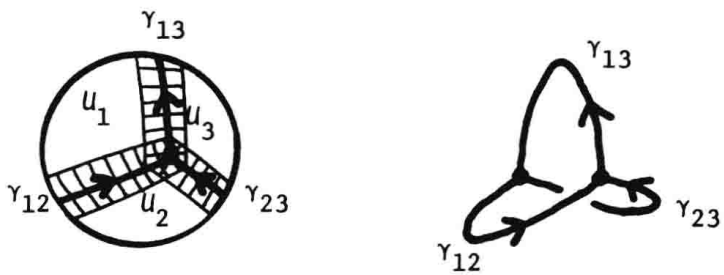
But we need not do so if we prefer not to. We have a generalization of the concept of a twistor function, namely as a collection of functions on portions of twistor space, defining a 1-cocycle with respect to some covering. We can actually use such a cocycle directly, obtaining the required space-time field by means of a branched contour integral. I shall just illustrate this with an example. Suppose that X is covered by 3 open sets U_1, U_2, U_3, where on a projective line L in X we get the picture:

 On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91



The 1-cocycle β is (f_{12}, f_{13}, f_{23}). To get the field $\phi\ldots.$ we perform the sum of three contour integrals with common end-points (in $U_1 \cap U_2 \cap U_3$):

$$(2\pi i)^2 \phi\ldots. = \int_{\gamma_{12}} \ldots. f_{12} d^2Z + \int_{\gamma_{13}} \ldots. f_{13} d^2Z + \int_{\gamma_{23}} \ldots. f_{23} d^2Z$$



It is a simple matter to check that the cocycle condition $f_{12} - f_{13} + f_{23} = 0$ (in $U_1 \cap U_2 \cap U_3$) ensures that the contour's end-points can be moved without affecting the result. This easily generalizes to coverings with N open sets.

On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978)*. Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

Some applications (where knowledge of sheaf cohomology theory would be helpful).

(a) Charge integrality in the "twisted photon" description as opposed to vanishing charge in the description

$$\phi_{AB} = \frac{1}{(2\pi i)^2} \oint \frac{\partial}{\partial \omega^A} \frac{\partial}{\partial \omega^B} f(Z^\alpha) d^2\pi$$

(f homogeneous of degree 0).

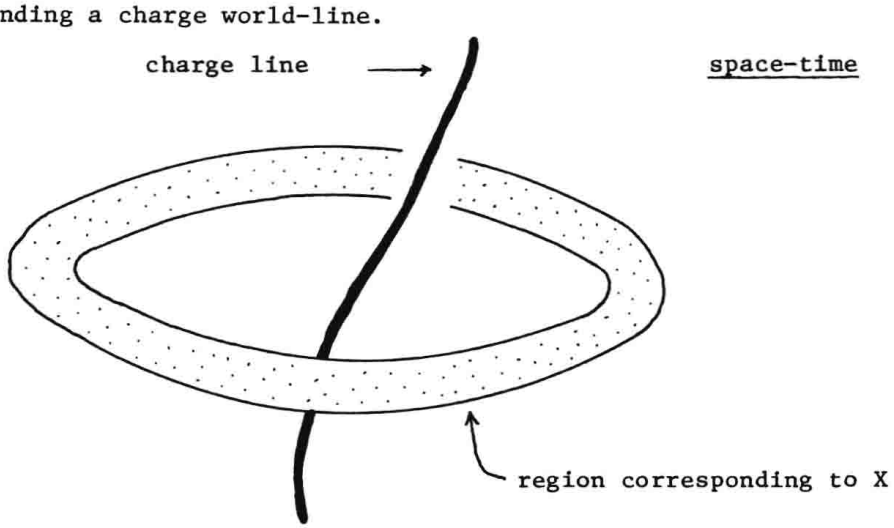
We have [3] the exact sequence

$$0 \to \mathbb{Z} \xrightarrow{\times 2\pi i} \mathcal{O} \xrightarrow{\exp} \mathcal{O}^* \to 0,$$

where $\mathcal{O}^*$ refers to non-zero holomorphic functions taken multiplicatively, from which we derive [3,4,5] the long exact sequence:

$$\ldots \to \underset{\substack{\uparrow \\ \text{ordinary integer 1st cohomology}}}{H^1(X,\mathbb{Z})} \to \underset{\substack{\uparrow \\ \phi_{\ldots} = \oint \frac{\partial}{\partial \omega} \frac{\partial}{\partial \omega} f}}{H^1(X,\mathcal{O})} \to \underset{\substack{\uparrow \\ \text{"twisted photon"}}}{H^1(X,\mathcal{O}^*)} \to \underset{\substack{\uparrow \\ \text{ordinary integer 2nd cohomology}}}{H^2(X,\mathbb{Z})} \to \ldots$$

Choose X as a region in $\mathbb{P}\mathbb{T}$ corresponding to a small space-time tube surrounding a charge world-line.



 On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

Then the topology of X is $S^2 \times S^2 \times R^2$, so $H^1(X,\mathbb{Z}) = 0$, and $H^2(X,\mathbb{Z}) \tilde{=} \mathbb{Z} \oplus \mathbb{Z}$. Thus the space $H^1(X,\mathcal{O}^*)$ effectively contains $H^1(X,\mathcal{O})$, and is strictly larger than $H^1(X,\mathcal{O})$ if the map to $\mathbb{Z} \oplus \mathbb{Z}$ is not simply to the zero element. The image in the first $\mathbb{Z}$ is always zero, but the image in the second $\mathbb{Z}$ is the value of the charge. (This much follows, for example, from examination of reference [6].) From this we see that the $H^1(X,\mathcal{O}^*)$ description only works if the charge has integer value (i.e. lies in $\mathbb{Z}$) whereas the $H^1(X,\mathcal{O})$ description only works if the charge value is zero. (In contrast to this, it may be remarked that there is no restriction on the charge value if the -4-homogeneity $f(Z^\alpha)$ description

$$\tilde{\phi}_{A'B'} = \frac{1}{(2\pi i)^2} \oint \pi_{A'} \pi_{B'} f(Z^\alpha) d^2\pi$$

is used, this corresponding to $H^1(X,\mathcal{O}(-4))$.)

(b) How to grow (or at least half-grow) new twistors from old.

There is another way of representing elements of $H^p(X,\mathcal{O})$, namely by taking the additive group of all $\bar{\partial}$-closed (0,p)-forms modulo the additive group of all $\bar{\partial}$-exact (0,p)-forms [3,5]. By a (q,p)-form is meant a (q+p)-form on X (not generally holomorphic) having q holomorphic differentials and p anti-holomorphic differentials:

$$\alpha_{i_1 \cdots i_q \, j_1 \cdots j_p} \, dz^{i_1} \wedge \cdots \wedge dz^{i_q} \wedge d\bar{z}^{j_1} \wedge \cdots \wedge d\bar{z}^{j_p} = \alpha.$$

Then $\bar{\partial}\alpha = \frac{\partial \alpha \ldots}{\partial \bar{z}^k} \, d\bar{z}^k \wedge dz^{i_1} \wedge \cdots \wedge d\bar{z}^{j_p}$ is a $\bar{\partial}$-exact (q,p+1)-form whose vanishing is the condition for α to be $\bar{\partial}$-closed. (We have $\bar{\partial}^2 = 0$.) Note that the $\bar{\partial}$-closed (0,0)-forms are holomorphic functions. We can also "twist" such forms or functions, as before.

Applying this to $H^1(X,\mathcal{O}(n))$ we obtain the result that the helicity $\frac{1}{2}(-n-2)$ massless fields can be described in terms of forms $g^\alpha(Z^\cdot,\bar{Z}.)d\bar{Z}_\alpha$ on X, modulo forms $\frac{\partial h(Z^\cdot,\bar{Z}.)}{\partial \bar{Z}_\alpha}\, d\bar{Z}_\alpha$, where $\frac{\partial g^\alpha}{\partial \bar{Z}_\beta} = \frac{\partial g^\beta}{\partial \bar{Z}_\alpha}$, the functions g^α being homogeneous of degree n in Z^α and of degree -1 in $\bar{Z}_\alpha$ (to balance the $d\bar{Z}_\alpha$). In order that the form $g^\alpha d\bar{Z}_\alpha$ be actually defined on the projective space (appropriately twisted) we require also $g^\alpha \bar{Z}_\alpha = 0$. The massless field integral now becomes

$$\phi_{A\ldots D} = \frac{1}{(2\pi i)^2}\oint \frac{\partial}{\partial \omega^A}\cdots\frac{\partial}{\partial \omega^D}\, g^\gamma(Z^\cdot,\bar{Z}.)d\bar{Z}_\gamma \wedge (I_{\alpha\beta}Z^\alpha dZ^\beta)$$

or

$$\phi_{A'\ldots D'} = \frac{1}{(2\pi i)^2}\oint \pi_{A'}\cdots\pi_{D'}\, g^\gamma(Z^\cdot,\bar{Z}.)d\bar{Z}_\gamma \wedge (I_{\alpha\beta}Z^\alpha dZ^\beta).$$

(See also N. M. J. Woodhouse's article in Twistor Newsletter 2 for further details [7].)

There is presumably much freedom in the choice of g^γ. I shall suppose that its singularities can be arranged so that the contour in the above integrals (which is initially S^2, i.e., over the entire projective line corresponding to the space-time point in question) may be deformed, freeing $\bar{Z}_\alpha$ from Z^α in the process, until W_α ($=\tilde{Z}_\alpha$ = "freed" $\bar{Z}_\alpha$) passes through the line I. Writing $W_\alpha = I_{\alpha\beta}X^\beta$, we get

$$g^\gamma d\bar{Z}_\gamma \wedge I_{\alpha\beta}Z^\alpha dZ^\beta = g^\rho I_{\rho\sigma}dX^\sigma \wedge Z^\alpha I_{\alpha\beta}dZ^\beta.$$

The function g^α can be taken to be holomorphic in Z^α and W_α, with $W_\alpha g^\alpha = 0$ and $\frac{\partial g^\alpha}{\partial W_\beta} - \frac{\partial g^\beta}{\partial W_\alpha} = 0$. Restricting to $W_\alpha = I_{\alpha\beta}X^\beta$ we get $I_{\alpha\beta}g^\alpha X^\beta = 0$ whence

 On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

$g^{\alpha}I_{\alpha\beta} = qX^{\alpha}I_{\alpha\beta}$ for some $q(Z^{\alpha}, I_{\alpha\beta}X^{\beta})$ homogeneous of degrees (n,-2). Thus, the massless field integral can be written

$$\phi_{\dots\dots} = \frac{1}{(2\pi i)^2}\oint_{S^2}\left\{\begin{array}{l}\frac{\partial}{\partial\omega}\dots\\ \text{or}\\ \pi\dots\end{array}\right\} qX^{\alpha}I_{\alpha\beta}dX^{\beta}\wedge Z^{\rho}I_{\rho\sigma}dZ^{\sigma},$$

which is the standard form for a 2-twistor integral. Thus we have conjured a new twistor out of thin air! This new twistor X^{α} is still only half-grown, though, since it appears as $I_{\alpha\beta}X^{\beta}$. This can be stated as $I^{\alpha\beta}\frac{\partial q}{\partial X^{\beta}} = 0$ which implies the massless relation $I_{\rho\sigma}Z^{\rho}X^{\sigma}I^{\alpha\beta}\frac{\partial^2 q}{\partial Z^{\alpha}\partial X^{\beta}} = 0$. In fact, the $\bar{\partial}$-closed property $\frac{\partial g^{\alpha}}{\partial W_{\beta}} - \frac{\partial g^{\beta}}{\partial W_{\alpha}} = 0$ now disappears on $W_{\alpha} = I_{\alpha\beta}X^{\beta}$, $g^{\alpha} \mapsto g^{\beta}I_{\beta\alpha}$, as it is incorporated in $I^{\alpha\beta}\frac{\partial q}{\partial X^{\beta}} = 0$ and $X^{\alpha}\frac{\partial q}{\partial X^{\alpha}} = -2q$ (homogeneity relation).

Acknowledgment

I am grateful to George Sparling, Andrew Hodges, and Nick Woodhouse for many helpful discussions and, most particularly, to Michael Atiyah for explaining sheaf cohomology theory so clearly to us and for many insightful remarks.

2. THE BACK-HANDED PHOTON (or, to a cricketer, THE GOOGLY PHOTON) *

Current twistor dogma presents the following picture of things:

(i) The wave function of a particle or system of particles is to be given by a "twistor function" $\psi(X^{\alpha},\dots,Z^{\alpha})$, which is holomorphic in all arguments $X^{\alpha},\dots,Z^{\alpha}$ --possibly with the modification that some (or all) of these should be dual twistors $W_{\alpha},\dots$ instead (or, conceivably, Hughston-type multi-twistors $X^{\alpha\beta\gamma},\dots$, etc.).

*Taken from Twistor Newsletter 3 (December, 1976).

(ii) More correctly, ψ is (presumably) an element of a sheaf cohomology group, say $\psi \in H^k(Q,\mathcal{O})$, for some suitable open set (or closure of an open set) $Q \subset \mathbb{T} \times \mathbb{T} \times \ldots \times \mathbb{T}$ say ($\mathbb{T}$ = twistor space), $\mathcal{O}$ being the sheaf of (germs of) holomorphic functions on Q (see previous section).

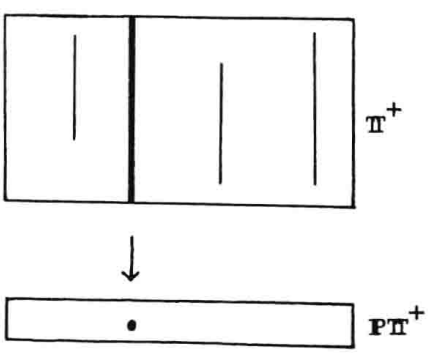
(iii) When a particle interacts it does so by deforming the structure of twistor space, or of $Q \subset \mathbb{T} \times \ldots \times \mathbb{T}$, whereby ψ now plays an active role in determining the precise deformation in question, thence affecting the behavior of other twistor functions defined on Q.

The above is vague in various respects. But something like (iii) is strongly suggested by two examples: the non-linear graviton [8] and the twisted photon [9]. (In fact, for gravitons, (iii) is strictly true only in the weak-field limit--unless some kind of non-linear sheaf theory can be evoked.) In the standard procedures for performing deformations (infinitesimal ones at least) it is only H^1's that are involved. Sparling has suggested that new procedures for performing deformations may be needed in order to be able to use H^k's ($k > 1$) in an active way. For the moment, I prefer the more modest idea that single particles are to be described by H^1's, these being the things that interact directly with other particles. H^2's, H^3's, etc. would play roles in describing many-particle states, with elements of suitable $H^k(Q,\mathcal{O})$ groups defining k-particle states. This could supply a possible answer to a question posed to me by Feynman: how do you know whether $\psi(X^\alpha,Y^\alpha,Z^\alpha)$ describes a hadron, or three massless particles, or a lepton and a massless particle? Elements of $H^1(Q,\mathcal{O})$, $H^3(Q,\mathcal{O})$, and $H^2(Q,\mathcal{O})$ are really quite different kinds of animals, after all! But the detailed implementation of this idea has proved elusive--not the least problem being to understand "twistorially" why the spin-statistics relation should hold for many-particle states.

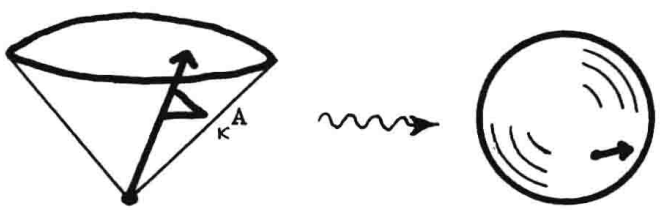
 On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

Perhaps the most primitive problem, in trying to push forward with (iii), lies in the fact that a twisted photon is only <u>half</u> a photon, namely the left-handed half. Thus, the Ward(-Sparling) construction gives a deformation of $Q = \mathbb{T}^+$ starting from a twistor function $f(Z^\alpha)$, homogeneous of degree zero ($f \in H^1(\mathbb{T}^+, \mathcal{O})$ or $f \in H^1(\mathbb{P}\,\mathbb{T}^+, \mathcal{O}(0))$), to give a <u>left</u>-handed photon. Of course a right-handed photon can be produced by using $\tilde{f}(W_\alpha)$, of degree zero. But such would be to defeat the purpose of the economy of the twistor description (i). Simply changing the helicity quantum number (or any other quantum number) should not involve us in changing the <u>space</u> Q over which the twistor function is defined. Thus maintaining the general programme (i) - (iii) seems to lead us to the view that some form of deformation of $\mathbb{T}^+$ <u>must</u> be possible, which effectively encodes the information provided by an element $f \in H^1(\mathbb{T}^+, \mathcal{O})$, when f is homogeneous of degree -4 (i.e., in effect, $f \in H^1(\mathbb{P}\,\mathbb{T}^+, \mathcal{O}(-4))$. My suggestion for this, which relates closely to an earlier proposal due to Sparling and Ward [10], is as follows.

Consider, first, the standard Ward twisted photon. This is obtained by deforming the bundle (see figure) where the base space remains the undeformed



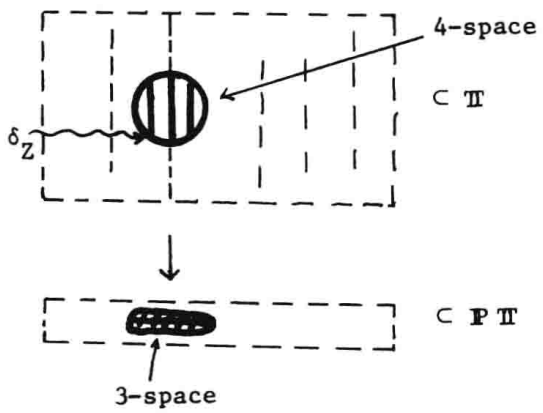
$\mathbb{P}\mathbb{T}^+$ and the fibre $\mathbb{C} - \{0\}$. Furthermore, the Euler operator $Z^\alpha \frac{\partial}{\partial Z^\alpha} =: T_Z$ remains globally defined. However, with non-trivial twisting, the forms $\mathbb{D}Z := \frac{1}{24}\,\varepsilon_{\alpha\beta\gamma\delta}\, dZ^\alpha \wedge dZ^\beta \wedge dZ^\gamma \wedge dZ^\delta$ and $\mathcal{D}Z := \frac{1}{6}\,\varepsilon_{\alpha\beta\gamma\delta}\, Z^\alpha dZ^\beta \wedge dZ^\gamma \wedge dZ^\delta$ are <u>not</u> well defined; indeed, globally defined analogues of these forms do not exist at all on a non-trivially twisted Ward photon. There is a good reason for this. In the case of flat $\mathbb{T}$, there is a canonical way of representing a twistor Z^α (up to the fourfold ambiguity $\pm Z^\alpha$, $\pm iZ^\alpha$) in terms of $\mathbb{P}\mathbb{T}$, namely as the projective twistor $\mathbb{Z}$ in $\mathbb{P}\mathbb{T}$, together with a <u>3-form at</u> $\mathbb{Z}$ in $\mathbb{P}\mathbb{T}$. This 3-form lifts into $\mathbb{T}$ along the fibre over $\mathbb{Z}$; where it agrees with $\mathcal{D}Z$ defines us the point Z^α (or iZ^α, or $-Z^\alpha$, or $-iZ^\alpha$). Thus, if $\mathcal{D}Z$ is canonically known in the bundle, then the complete bundle structure is determined uniquely in terms of $\mathbb{P}\mathbb{T}$; therefore no twisting in the "photon" can occur. (This construction is the analogue of how one defines a spinor κ^A (up to sign) in terms of the celestial sphere.)



Furthermore, if a $\mathcal{D}Z$ exists globally in a bundle over $\mathbb{P}\mathbb{T}^+$, then it must be unique up to an overall constant factor.

 On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

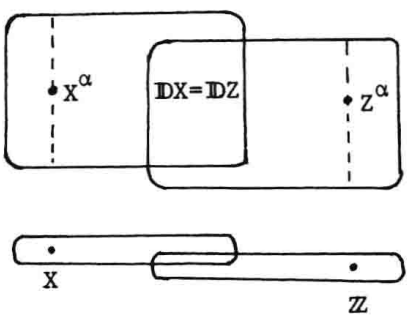
Suppose, generally, we have a bundle (or holomorphic fibration) just locally, which is just a complex 4-space over a complex 3-space with fibre a complex 1-space. Then all we know locally in the 4-space is a direction



field δ_Z (1-foliation). To know $\mathcal{D}Z$ in the 4-space would be to know rather more structure than δ_Z. Being a 3-form, $\mathcal{D}Z$ is orthogonal to (i.e., annihilates) precisely one complex direction and so serves to define δ_Z. (I shall always assume $\mathcal{D}Z$ to be restricted to be orthogonal to δ_Z, so as to qualify to be a "$\mathcal{D}Z$.") Now $\mathcal{D}Z$ also defines a volume 4-form $\mathbb{D}Z$ by $d\mathcal{D}Z = 4\mathbb{D}Z$. Furthermore $\mathcal{D}Z$ and $\mathbb{D}Z$ together define the Euler operator T_Z, roughly speaking by "$T_Z = \mathcal{D}Z \div \mathbb{D}Z$," or more precisely by $\mathcal{D}Z = T_Z \lrcorner \mathbb{D}Z$. Conversely, this relation shows us that $\mathcal{D}Z$ is determined by the pair $(T_Z, \mathbb{D}Z)$. To know one or the other of T_Z, $\mathbb{D}Z$ is, by itself, not sufficient to determine $\mathcal{D}Z$, but the two together are equivalent to $\mathcal{D}Z$. Furthermore, assuming that T_Z is to point along δ_Z (where δ_Z is given), there is precisely as much information in T_Z locally as there is in $\mathbb{D}Z$. Each provides us with a kind of local scaling, but it is a "homogeneity degree 4" scaling for $\mathbb{D}Z$

and a "homogeneity degree 0" scaling for T_Z. In a sense $\mathbb{D}Z$ and T_Z seem to be sorts of duals to one another.

We can regard the Ward photon [9] as arising when we retain only the T_Z scaling and throw out $\mathbb{D}Z$. Let us try to do the "dual" thing and retain $\mathbb{D}Z$ while throwing out T_Z. I shall proceed in a fairly explicit way, assuming that two "coordinate" patches are given, where a standard twistor description is given in each patch, with X^α on the left and Z^α on the right. I am



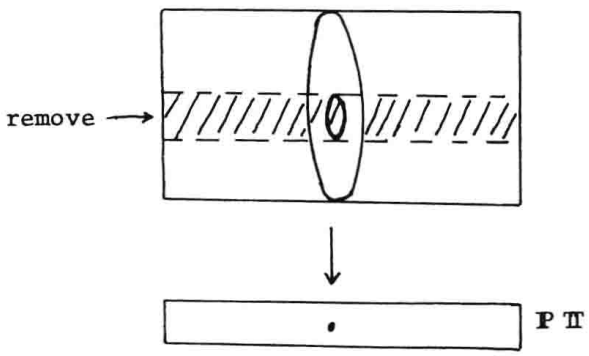
assuming there is no monkey-business in the base space, so $X^\alpha \propto Z^\alpha$ may be assumed on the overlap region. (The fibres in each half are given when X^α or Z^α is held constant up to proportionality.) The hypothesis is now $\mathbb{D}X = \mathbb{D}Z$ on the overlap (instead of Ward's $T_X = T_Z$), i.e., $d\mathcal{D}X = d\mathcal{D}Z$, i.e., $d\{\mathcal{D}X-\mathcal{D}Z\} = 0$. But $\delta_X = \delta_Z$, so $\mathcal{D}X \propto \mathcal{D}Z$. Put $\mathcal{D}X = \{1 + f(Z^\alpha)\}\mathcal{D}Z$; then $d\{f(Z^\alpha)\mathcal{D}Z\} = 0$, which holds iff $f(Z^\alpha)$ is homogeneous of degree -4. The transition relation is then

$$X^\alpha = \{1 + f_{-4}(Z^\alpha)\}^{\frac{1}{4}} Z^\alpha. \qquad \text{(A)}$$

 On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

This appears to be a particular case of a Sparling-Ward [10] construction given earlier for massless fields of any helicity.

We run into trouble owing to branch points arising from the fourth root unless we exclude an extended region about the origin of each fibre. But things are okay near ∞ on the fibres. Thus, I envisage the fibres as being like $\mathbb{C}$ with some bounded (probably connected) region removed (i.e.,



biholomorphic to $\mathbb{C} - \{z : |z| < 1\}$). Note that T_Z is not preserved by the patching, but related by $T_X = (1 + f(Z^\alpha))T_Z$ (which does strange things to the notion of a homogeneous function).

The intention is to regard $f(Z^\alpha)$ as a twistor function for a right-handed photon. But, in accordance with (ii), such $f(Z^\alpha)$ is really describing an element of a sheaf cohomology group. Thus we need to check that (A) is appropriately cohomological. Suppose we have a covering of the base space (say $\mathbb{P}\mathbb{T}^+$) by a number of open sets U_i. Then on each $U_i \cap U_j$ we need $f_{ij} = -f_{ji}$ such that on each $U_i \cap U_j \cap U_k$, $f_{ik} = f_{ij} + f_{jk}$ (cocycle condition). If we use standard twistor "coordinates" Z^α for the entire base space, then the f_{ij} will simply be homogeneous functions of Z^α, of

degree -4. Thus, $f_{ij}(Z^\alpha) = -f_{ji}(Z^\alpha)$, $f_{ik}(Z^\alpha) = f_{ij}(Z^\alpha) + f_{jk}(Z^\alpha)$. The cocycle defining the right-handed photon can be taken to be just this collection of homogeneous functions. We piece together our "bundle" by taking "coordinates" Z^α_i for the portion lying above U_i, where on each overlap (say, the portion of the bundle lying above $U_i \cap U_j$) we have $Z^\alpha_i = \{1 + f_{ij}(Z^\beta_j)\}^{\frac{1}{4}} Z^\alpha_j$. We must first check the consistency of this with $f_{ij} = -f_{ji}$. Avoiding messy indices we have (A) : $X^\alpha = \{1 + f(Z^\beta)\}^{\frac{1}{4}} Z^\alpha$ and $Z^\alpha = \{1 + g(X^\beta)\}^{\frac{1}{4}} X^\alpha$, and we need to check that $g = -f$. Now $Z^\alpha = \{1 + g(X^\beta)\}^{\frac{1}{4}}\{1 + f(Z^\beta)\}^{\frac{1}{4}} Z^\alpha$, so we require $(1 + g(X^\beta))(1 + f(Z^\beta)) = 1$; that is

$$1 = (1 + g(\{1 + f(Z^\beta)\}^{\frac{1}{4}} Z^\sigma))\,(1 + f(Z^\rho))$$

$$= (1 + \frac{g(Z^\sigma)}{1 + f(Z^\beta)})(1 + f(Z^\rho))$$

$$= (1 + f(Z^\rho) + g(Z^\rho))$$

(because of the -4 homogeneity of g) so $g(Z^\rho) = -f(Z^\rho)$ as required. Next we check the compatibility of (A) with $f_{ik} = f_{ij} + f_{jk}$. Again avoiding indices we have, on the triple intersection region above $U_i \cap U_j \cap U_k$ relations

$$Y^\alpha = \{1 + p(X^\rho)\}^{\frac{1}{4}} X^\alpha;\quad X^\rho = \{1 + f(Z^\sigma)\}^{\frac{1}{4}} Z^\rho;\quad Y^\alpha = \{1 + q(Z^\sigma)\}^{\frac{1}{4}} Z^\alpha$$

For consistency we require

$$\{1 + p(\{1 + f(Z^\sigma)\}^{\frac{1}{4}} Z^\rho\}\,\{1 + f(Z^\beta)\} = 1 + q(Z^\sigma)$$

i.e. $$\{1 + \frac{p(Z^\rho)}{1 + f(Z^\rho)}\}\{1 + f(Z^\rho)\} = 1 + q(Z^\rho)$$

i.e. $q(Z^\rho) = p(Z^\rho) + f(Z^\rho)$ as required. We need also to show that if the cocycle is a coboundary, i.e., if $f_{ij} = f_j - f_i$ for each i,j, then the bundle

 On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

is the same as that for flat twistor space (f_i being holomorphic throughout U_i). For this we set $Z^\alpha = \{1 + f_i(Z_i^\beta)\}^{\frac{1}{4}} Z_i^\alpha$ and find that the Z^α, so constructed, is defined globally over the whole bundle. The compatibility over each $U_i \cap U_j$ follows by a calculation which is basically the same as the ones given above. Finally we need the fact that if the bundle is the same as that for flat twistor space (where here and above "the same" must be suitably interpreted in terms of "analytically extendible to"--owing to the gaping holes in the fibres), then the $\{f_{ij}\}$-cocycle is a coboundary. To prove this, we simply reverse the above argument. The flatness implies the existence of a global Z^α for the bundle. The local Z_i^α for each patch (above U_i) must be related to Z^α by a formula $Z^\alpha = \{1 + f_i(Z_i^\beta)\}^{\frac{1}{4}} Z_i^\alpha$ from which follows (by the reverse of the above calculation) $f_{ij} = f_j - f_i$ above $U_i \cap U_j$.

The argument just given effectively shows that although (A) appears to be a highly non-linear relation, the system of bundles so constructed has nevertheless a linear structure (given by simply adding the sheaf cohomology group elements). This may be contrasted with Sparling's method of patching together non-linear gravitons:

$$X^\alpha = \exp\{I^{\rho\sigma} \frac{\partial f(Z^\tau)}{\partial Z^\rho} \frac{\partial}{\partial Z^\sigma}\} Z^\alpha.$$

For this, the cocycle condition fails. This is a manifestation of the very non-linearity of the non-linear graviton. On the other hand, the "4" in the above construction is not essential for linearity and could be replaced by other powers; i.e., f of homogeneity -n, with $X^\alpha = \{1 + f(Z^\beta)\}^{1/n} Z^\alpha$. When n=1, "0", -1, -2, -3,... this is implicit in the Sparling-Ward construction. However, when $n < 0$ the global structure of the fibres would have to be something different (because $(1 + z^{-n})^{1/n} z \to 1$ as $|z| \to \infty$ if $n < 0$, whereas

$(1 + z^{-n})^{1/n} z \sim z$ as $|z| \to \infty$ if $n > 0$). In any case, the motivation from "preservation of $\mathbb{DZ}$" exists only when $n = 4$.

A direct construction of the right-handed Maxwell field $\tilde{\phi}_{A'B'}$ (satisfying $\nabla^{AA'} \tilde{\phi}_{A'B'} = 0$) from the bundle structure has not yet emerged. But the most serious problem is that of fitting the right-handed and left-handed ways of deforming $\mathbb{T}^+$ together into one bundle...? One possibility for doing this is a method suggested by Hughston [12]. However it is unclear, as yet, how to extract left- and right-handed Maxwell fields which do not interact with one another. If this problem can be resolved, then among other things, it might suggest an analogous approach to an ambidextrous graviton!

3. REMARKS CONCERNING THE AMBIDEXTROUS PHOTON *

The information of a left-handed Maxwell field can be stored [9] in a holomorphic line bundle T over a suitable portion R of $\mathbb{P}\mathbb{T}$, the Euler operator T being well-defined on T and pointing along the fibres of T. Likewise, the information of a right-handed Maxwell field can be stored in a 1-manifold-fibration T over R, but where now it is the holomorphic volume 4-form $\Delta(=\mathbb{DZ})$ which is well-defined on T instead of T. (See the previous section and also reference [9].) In reference [12], Hughston suggested putting the right-handed construction on top of the left-handed one, but it was not clear that this preserved any particular structure on T. In fact it does, namely the operator I shall call DIV. This is a map from vector fields V, pointing along the fibration, to scalars, satisfying $\mathrm{DIV}(\lambda V) = \lambda \mathrm{DIV}(V) + V(\lambda)$, where λ is a scalar field. It differs from an ordinary div operation only in that it need not act on any vector field which does not point along the fibration. Clearly, if Δ exists, then DIV exists, defined by $\mathrm{DIV}(V) = \mathrm{div}_{\Delta} V$, where

*Taken from Twistor Newsletter 5 (July, 1977).

 On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

div_Δ is the ordinary divergence based on Δ (i.e., use Δ to convert V into a 3-form $V \lrcorner \Delta$, take the exterior derivative of $V \lrcorner \Delta$, and then divide out by Δ to obtain a scalar). Also, if T exists, then DIV exists, defined by $T \times \mathrm{DIV}(V) = [T,V] + 4V$.

Let us now suppose that the 4-dimensional space T, which is a fibration over R (all holomorphic), possesses just the structure DIV. Cover R by open sets $\{U_i\}$ and lift these to give a covering $\{\mathcal{U}_i\}$ of T. Define a Δ_i and an T_i in each $\mathcal{U}_i$ which are compatible, in the sense above, with DIV. These compatibility requirements can be expressed as

$$\mathrm{DIV}(T_i) = 4 \quad \text{and} \quad \pounds_{T_i} \Delta_i = 4\Delta_i$$

(the latter being equivalent to $\mathrm{div}_{\Delta_i} T_i = 4$). Now define $f_{ij} = \log(\Delta_j/\Delta_i)$. We have $\pounds_{T_i} f_{ij} = 0$ and $f_{ij} - f_{ik} + f_{jk} = 0$ on $\mathcal{U}_i \cap \mathcal{U}_j \cap \mathcal{U}_k$, so f_{ij} defines a 1-cocycle on R. This is the left-handed (homogeneous of degree 0) "part" of the photon. Whenever this cohomology class vanishes, Δ exists globally. Suppose, next, that Δ <u>is</u> global. Define $F_{ij} = (T_j - T_i) \lrcorner \Delta$. We find $\pounds_{T_i} F_{ij} = 0$ and $F_{ij} - F_{ik} + F_{jk} = 0$ on $\mathcal{U}_i \cap \mathcal{U}_j \cap \mathcal{U}_k$, so F_{ij} defines a 1-cocycle 3-form on R. This is the right-handed (back-handed) (homogeneous of degree -4) "part" of the photon.

In fact, this is exactly what we had before (in Twistor Newsletter 3 and [10]) although phrased somewhat differently, so that we don't even notice the $(\ldots)^{\frac{1}{4}}$ unpleasantness! (Incidentally, another way of disguising this is to write the $\hat{Z} = (1 + 4g)^{\frac{1}{4}} Z$ as $\hat{Z} = e^{gT} Z$, where g(Z) is homogeneous of degree -4. The power series <u>do</u> check!)

The trouble, of course, is the same as before: if there <u>is</u> a left-handed part (Δ <u>not</u> global) then there seems to be no way of extracting an

$H^1(R,\mathcal{O}(-4))$ element to measure the non-globalness of T (without bringing in more structure, that is). This is the old problem--the right-handed part appears to be charged with respect to the left-handed part. To decouple the two parts, it seems that strict conformal invariance must be broken (e.g. $F_{ij} = (T_j - T_i) \lrcorner \Delta_i$ or $F_{ij} = (T_j - T_i) \lrcorner (\Delta_j - \Delta_i)$ won't do because the cocycle condition fails). Thus we must bring in I (the "infinity twistor"). This works in principle, but is, as yet, inelegant.

4. MASSLESS FIELDS AND SHEAF COHOMOLOGY*

In the first of these four Twistor Newsletter articles I indicated that twistor functions are really to be regarded as providing representative cocycles for sheaf cohomology, so the twistorial representation of a massless field of "helicity $\frac{n}{2}$" is an element of $H^1(X,\mathcal{O}(-n-2))$, where X is an (open) region in $\mathbb{P}\mathbb{T}$ (say $\mathbb{P}\mathbb{T}^+$) swept out by the projective lines which correspond to the points of a suitable given (open) region Y in $\mathbb{CM}$ (say $\mathbb{CM}^+$) where the field is defined. Woodhouse [7] showed how the space-time field could be obtained from the $H^1(...)$ element, when the $H^1(...)$ element is defined by means of Dolbeault [3,4] cohomology (i.e., "anti-holomorphic" forms). A method due to Ward shows how to obtain the field from a Čech cohomology cocycle f_{ij} using an extension of Sparling's "splitting" formula (cf. Ward [9]). Ward's method essentially replaces the "branched contour integral" described in the first section (cf. also Ward [10]). Basically, Ward's method is as follows:

Let $f_{ij}(Z^\alpha)$ be homogeneous of degree -n-2 (with $n \geq 0$) and defined on $U_i \cap U_j$ (with $\{U_i\}$ a covering of $X \subset \mathbb{P}\mathbb{T}$), $Z^\alpha = (\omega^A, \pi_{A'})$, $\omega^A = ix^{AA'}\pi_{A'}$, etc. Then $\pi_{A'} \ldots \pi_{E'} f_{ij}(Z^\alpha)$ (where π appears n+1 times) is homogeneous of

*Taken (with minor modifications) from Twistor Newsletter 5 (July, 1977).

 On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

degree -1 in π, so ("splitting") we get (ρ_i meaning "restriction," adopting a notational device due to Hughston, where square brackets denote skew-symmetrization):

$$\pi_{A'}\ldots\pi_{E'}f_{ij}(Z^{\alpha}) = \rho_{[i}\psi_{j]A'\ldots E'}(x^{i},\pi_{R'})$$

whence

$$\phi_{jA'\ldots D'}(x,\pi) := \pi^{E'}\psi_{jA'\ldots D'E'}$$

satisfies $\rho_{[i}\phi_{j]A'\ldots D'} = 0$. Thus, due to the sheaf property $\rho_{[i}F_{j]} = 0 \Rightarrow F_j = \rho_j G$, we have

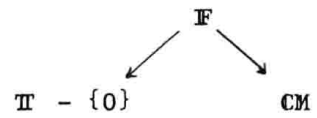
$$\phi_{jA'\ldots D'}(x,\pi) = \rho_j\phi_{A'\ldots D'}$$

and, since $\phi_{A'\ldots D'}$ is homogeneous of degree 0 in π (and global in π), it must be a function of x only. Furthermore, $\nabla^{AA'}\phi_{A'B'\ldots D'} = 0$ readily follows. The case $n < 0$ is somewhat similar.

All this goes one way only: from twistor function to space-time field. But using the method of exact homology sequences we can readily extend this to obtain (implicitly) a version of the "inverse twistor function," which goes from space-time field to twistor function. (Compare also the Bramson-Sparling-Penrose method outlined in [1].) The following was inspired to some extent also by Hughston [17], cf. also Lerner [18].

Let $\mathbb{F}$ be the (dual) primed spin-vector bundle (excluding the section of zero spin vectors) over $\mathbb{CM}$. Thus, a point of $\mathbb{F}$ can be labelled $(x^{a},\pi_{A'})$ (with $\pi_{A'} \neq 0$)--except for points at infinity on $\mathbb{CM}$--with x^{a} complex. $\mathbb{F}$ is also a bundle over $\mathbb{PT} - \{0\}$, the fibre over Z^{α} being the set of linear 2-spaces containing Z^{α}.

On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978)*. Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91



Consider the following exact sequence of sheaves over $\mathbb{F}$:

$$0 \longrightarrow T \xrightarrow{\times \pi_{A'} \cdots \pi_{E'}} Z'_{n+1} \xrightarrow{\times \pi^{E'}} Z'_{n} \longrightarrow 0. \qquad \text{(B)}$$

Here Z'_n stands for the sheaf of n-index symmetric spinor fields $\phi_{A'\ldots D'}(x^i, \pi_{Q'})$ which are holomorphic in x^i and $\pi_{Q'}$, and satisfy the massless free-field equations $\nabla^{AA'}\phi_{A'B'\ldots D'} = 0$ ($\nabla_{AA'} = \frac{\partial}{\partial x^{AA'}}$, $\pi_{A'}$ being merely a passenger). The map $Z'_{n+1} \longrightarrow Z'_n$ is simply

$\psi_{A'\ldots D'E'} \longmapsto \phi_{A'\ldots D'} = \pi^{E'}\psi_{A'\ldots D'E'}$ and we must check that is is <u>onto</u>, i.e., that given $\phi\ldots$ we can always solve for $\psi\ldots$. One method is as follows: <u>First</u>, fix $\pi_{A'}$, and choose a primed spinor basis $o^{A'}, \iota^{A'}$ with $\pi^{A'} = o^{A'}$. We have $\psi_{0'\ldots0'0'} = \phi_{0'\ldots0'}$, $\psi_{1'0'\ldots0'0'} = \phi_{1'0'\ldots0'}$, $\cdots$, $\psi_{1'\ldots1'0'} = \phi_{1'\ldots1'}$, but $\psi_{1'\ldots1'1'}$ is fixed only by the massless field equations

$$-\frac{\partial}{\partial x_{01'}}\psi_{1'\ldots1'1'} = \frac{\partial}{\partial x_{00'}}\psi_{1'\ldots1'0'} \left(= \frac{\partial}{\partial x_{00'}}\phi_{1'\ldots1'}\right)$$

and

$$-\frac{\partial}{\partial x_{11'}}\psi_{1'\ldots1'1'} = \frac{\partial}{\partial x_{10'}}\psi_{1'\ldots1'0'} \left(= \frac{\partial}{\partial x_{10'}}\phi_{1'\ldots1'}\right),$$

the integrability conditions for which come simply from the massless equations on $\phi_{A'\ldots D'}$. We can ensure a unique solution locally in $\mathbb{CM}$ by choosing (say) $\psi_{1'\ldots1'1'} = 0$ on an initial 2-surface $x_{11'} = \text{constant}$, $x_{01'} = \text{constant}$. <u>Second</u>, allow $\pi^{A'}$ to vary locally in $\mathbb{C}^2 - \{0\}$; take $\iota^{A'}$ to

 On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

depend holomorphically on $\pi^{A'}$ ($= o^{A'}$) (say $\iota^{A'} \propto$ constant). The above prescription for finding $\psi_{A'\ldots D'E'}$ ensures that it is holomorphic in $\pi_{A'}$-- though the dependence on $\pi_{A'}$ may be complicated owing to the fact that the functional dependence of $\phi\ldots$ on the components $x_{\underline{A}0'}$ is π-dependent. (Incidentally, for simplicity, choose the unprimed basis constant.)

To see what $\mathcal{T}$ is we examine the kernel of the above map, i.e., find ψ's for which $\pi^{E'}\psi_{A'\ldots D'E'} = 0$ ($\pi^{E'} \neq 0$). We have (by standard lemmas) $\psi_{A'\ldots E'} = \pi_{A'}\cdots\pi_{E'}f$, and the massless equations on $\psi\ldots$ become $\pi^{A'}\nabla_{AA'}f = 0$. Thus the dependence of f on x^a is only through the quantity $\omega^A = ix^{AA'}\pi_{A'}$ (i.e., f is constant on twistor planes in $\mathbb{CM}$), so $f = f(\omega^A,\pi_{A'}) = f(Z^\alpha)$ is a twistor function. We see that $\mathcal{T}$ is the sheaf of twistor-type functions on $\mathbb{PT}$; equivalently, of holomorphic functions on $\mathbb{T} - \{0\}$.

The short exact sequence (B) gives rise to the long exact homology sequence:

$$\ldots \to H^0(Q,\mathcal{Z}'_{n+1}) \to H^0(Q,\mathcal{Z}'_n) \to H^1(Q,\mathcal{T}) \to H^1(Q,\mathcal{Z}'_{n+1}) \to \ldots \qquad (C)$$

Now, restrict $\mathcal{Z}'_n$ to be homogeneous of degree 0 in $\pi_{A'}$ (sheaf $\mathcal{Z}'_n(0)$); correspondingly, $\mathcal{Z}'_{n+1}$ to be homogeneous of degree -1 in $\pi_{A'}$ (sheaf $\mathcal{Z}_{n+1}(-1)$); so that $\mathcal{T}$ will be homogeneous of degree -n-2 in $\pi_{A'}$ (sheaf $\mathcal{T}(-n-2)$), i.e., of degree -n-2 in Z^α. Choose $Q \subset \mathbb{PT}$ to correspond to the region of $\mathbb{PT}$ lying above some $Y \subset \mathbb{CM}$ (so, for each $(x^a,\pi_{A'}) \in Q$ we have $(x^a,\hat{\pi}_{A'}) \in Q$ whenever $\hat{\pi}_{\underline{A}'} \in \mathbb{C}^2 - \{0\}$). Then since H^0 means "global sections" and since $\mathcal{Z}'_{n+1}$ is homogeneous of degree -1 in $\pi_{A'}$, we have

$$H^0(Q,\mathcal{Z}'_{n+1}(-1)) = 0.$$

To study $H^1(Q,Z'_{n+1})$ we can use, for example, a "resolution" of Z'_{n+1} (where dependence on $\pi_{A'}$ is irrelevant)

$$0 \to Z'_{n+1} \to F_{0,n+1} \to F_{1,n} \to F_{0,n-1} \to 0, \qquad (D)$$

this being exact, where $F_{p,q}$ is the sheaf of fields $\chi^{A\ldots C}_{D'\ldots F'}(x^i,\pi_{R'})$ with p symmetric unprimed indices and q symmetric primed ones (freely holomorphic), so $F_{p,q}$ is a coherent analytic sheaf, yielding $H^r(S,F_{p,q}) = 0$ whenever S is Stein and $r > 0$. The maps are:

$$\phi_{A'\ldots E'} \longmapsto \chi_{A'\ldots E'} = \phi_{A'\ldots E'},$$

$$\chi_{A'B'\ldots E'} \longmapsto \theta^{A}{}_{B'\ldots E'} = \nabla^{AA'}\chi_{A'B'\ldots E'},$$

$$\theta^{A}_{B'C'\ldots E'} \longmapsto \eta_{C'\ldots E'} = \nabla^{B'}_{A}\theta^{A}{}_{B'C'\ldots E'},$$

and exactness is not hard to verify. If n=0, the sequence terminates one step sooner (and if n=−1 we also have a short exact sequence

$$0 \to Z'_0 \xrightarrow{i} F_{0,0} \xrightarrow{\Box} F_{0,0} \to 0$$

where $\Box$ is the D'Alembertian). Note that (D) is conformally invariant, the conformal weights of $\phi\ldots$, $\chi\ldots$, $\theta\ldots$, $\eta\ldots$ being, respectively, −1, −1, −3, −4. The sequence (D) (for which $\pi_{A'}$ is irrelevant) is of interest in the study of massless fields quite independently of twistor theory. Here we just use it to derive the fact that

$$H^1(Q,Z'_{n+1}) = 0$$

whenever Y is Stein (as is the case if Y is the future- (or past-) tube in $\mathbb{CM}$, or if Y is a suitable "thickening" of a portion of the real space-time

 On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

M. (This follows because $H^1(Q,\mathcal{F}_{0,n+1}(-1))$ then vanishes from the homogeneity degree -1 in $\pi_{A'}$, the fact that $H^1(\mathbb{CP}_1,\mathcal{O}(-1)) = 0$ and suitable general theorems--for which thanks are due to R. O. Wells--and because $H^0(Q,\mathcal{F}_{0,n+1}(-1)/\mathcal{Z}'_n(-1))$ also vanishes--more obviously, because of -1 degree homogeneity.)

The upshot of this is that since the first and last terms in (C) both vanish (Y Stein), we have (for $n \geq 0$)

$$H^0(Q,\mathcal{Z}'_n(0)) \cong H^1(Q,\mathcal{T}(-n-2))$$

which, because globality in π on the left-hand-side implies constancy in π, almost establishes the required isomorphism between massless fields and H^1's of twistor functions. The remaining essential problem (the non-triviality and solution of which have been pointed out to me by M. Eastwood) is that all the sheaves refer to the space $\mathbb{F}$, so far, and some subtleties are involved in projecting the cohomology groups down to $\mathbb{T}$. The details of this and other matters are left to a proposed later paper to be written jointly with M. Eastwood and R. O. Wells, Jr. In fact the argument can be shown to work not only for the future tube (i.e., for $X = \mathbb{P}\mathbb{T}^+$) but also for suitable open regions in M. (Another subtlety has been glossed over in that the (x,π) description doesn't work for points at infinity. But this is unimportant because of conformal invariance and the fact that infinity can be transformed to somewhere safely inside the singular region of the field.)

To deal with the cases $n < 0$ we need a different exact sequence:

$$0 \to \mathcal{P}_{m-2} \xrightarrow{i} \mathcal{T} \xrightarrow{\frac{\partial}{\partial\omega^A}\cdots\frac{\partial}{\partial\omega^C}} \mathcal{Z}_{m-1} \xrightarrow{\pi^{D'}\nabla_{DD'}} \mathcal{Z}_m \to 0. \qquad (E)$$

On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

Here Z_m stands for the sheaf of unprimed massless holomorphic fields $\phi_{A\ldots D}(x^i,\pi_{R'})$ with m = -n indices, and P_{m-2} stands for twistor functions which are polynomials in ω^A of degree at most m-2 for each fixed $\pi_{A'}$. The fact that $Z_{m-1} \to Z_m$ is onto, is basically an argument given if reference [11]; the kernel is this: twistor functions $\psi_{A\ldots C}(Z^\alpha)$, the massless field equations on which are $\pi^{E'}\frac{\partial}{\partial\omega^{[D}}\psi_{A]\ldots C} = 0$, yielding $\psi_{A\ldots C} = \frac{\partial}{\partial\omega^A}\ldots\frac{\partial}{\partial\omega^C} f(Z)$ as required. The short exact sequence

$$0 \to T/P_{m-2} \to Z_{m-1} \to Z_m \to 0$$

provides the long exact sequence

$$H^0(Q,Z_{m-1}) \to H^0(Q,Z_m) \to H^1(Q,T/P_{m-2}) \to H^1(Q,Z_{m-1}). \qquad (F)$$

Choosing homogeneities -1, 0, m-2, -1, respectively, and using the same argument as above, we get (Y Stein):

$$H^0(Q,Z_m(0)) \cong H^1(Q,T(m-2)/P_{m-2}(m-2)). \qquad (G)$$

To deal with the extra complication of P_{m-2} we use $0 \to P_{m-2} \to T \to T/P_{m-2} \to 0$ to derive the exact sequence

$$H^1(Q,P_k(k)) \to H^1(Q,T(k)) \to H^1(Q,T(k)/P_k(k)) \to H^2(Q,P_k(k)) \qquad (H)$$

(where k = m-2). Now the sheaf $P_k(k)$ vanishes unless $k \geq 0$ (i.e., $m \geq 2$) and, using a suitable Kunneth formula, we derive the fact (since homogeneity in $\pi_{A'}$ for each x is non-negative) that we are concerned only with polynomial behavior in $\pi_{A'}$ for the two end terms in (H). Consequently these terms refer, in effect, to the "ordinary" cohomology of Q but with coefficients which are twistor polynomials $Z^\alpha\ldots Z^\gamma P_{\alpha\ldots\gamma}$, homogeneous of

 On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

degree k, i.e., equivalently, symmetric n-twistors $P_{\alpha\ldots\gamma}$. If Y describes a space-time region surrounding a "source tube," then we have $\frac{1}{6}(k+1)(k+2)(k+3) = \frac{1}{6}m(m^2-1)$ independent complex "charges," corresponding to the various independent components of $P_{\alpha\ldots\gamma}$, all of which have to vanish if the $H^1(Q,\mathcal{T})$ description is to work. This corresponds to the final map in (H) mapping to the zero element. (Note that m=2 for the anti-self-dual Maxwell case, which is consistent with a statement made in the first section of this article.) If Y is the forward tube $\mathbb{CM}^+$, then the first and last terms of (H) both vanish and the required isomorphism follows (but various details remain to be worked out).

Alternative approaches to the cases n < 0 can be given in which a potential rather than the field is used. This avoids having to bring in P_k and $P_{\alpha\ldots\gamma}$. (Work by Ward, Hughston and Eastwood.)

Further work is in progress.

Thanks are due to M. F. Atiyah, M. Eastwood, R. O. Wells, Jr. and R. S. Ward.

5. SOME REMARKS CONCERNING A GOOGLY* GRAVITON

The standard "non-linear graviton" construction [8] provides a means of coding a general anti-self-dual solution of Einstein's equations (or "right-flat space-time") into the structure of a deformed twistor space. A general self-dual solution (or "left-flat space-time") can be correspondingly coded into the structure of a deformed dual twistor space. However, it appears to be important, for the success of the twistor program as a whole, that the same twistor space be employable to encode the

*To those unfamiliar with the game of cricket, it should be explained that a "googly" is a ball bowled in such a way that it spins in a right-handed sense about its direction of motion even though, to the batsman, it would appear that the bowling action would be such as to impart a left-handed helicity.

information of <u>both</u> anti-self-dual and self-dual types of vacuum curvature. The hope would be that by combining both parts of the space-time (Weyl) curvature into the structure of <u>one</u> suitably deformed twistor space it might be possible to represent, twistorially, the general (analytic?) solution of Einstein's vacuum equations. I shall here be concerned mainly with the preliminary problem of attempting to encode the structure of a left-flat space-time into a deformed <u>twistor</u> space.

The essential difficulty in doing this appears to lie in the fact that the interpretation of a (complex) space-time point as a compact holomorphic curve in deformed projective twistor space, with null space-time separation being interpreted as intersection of the corresponding holomorphic curves, leads directly to the condition that many α-planes exist in the space-time and that consequently the space-time is right-conformally flat. What seems to be required is a different twistorial interpretation of a space-time point. Now recall [13] the simple exact sequence that expresses the essential Poincaré invariant structure of <u>flat</u> twistor space $\mathbb{T}$

$$0 \to S \xrightarrow{i} \mathbb{T} \xrightarrow{p} \tilde{\$}^* \to 0$$

where $\$$ is "unprimed" spin-space (the space of ω^A) and $\tilde{\$}^*$ is dual "primed" spin-space (the space of $\pi_{A'}$). The map i is $\omega^A \mapsto (\omega^A, 0)$ and p is $(\omega^A, \pi_{A'}) \mapsto \pi_{A'}$, with $Z^\alpha = (\omega^A, \pi_{A'})$. The standard twistorial representation of a (complex) space-time point may be thought of as a cross-section of $\mathbb{T}$, where $\mathbb{T}$ is regarded as a bundle over $\tilde{\$}^*$ with projection p, the cross-section being lined up along the direction of the Euler operator $T = Z^\alpha \partial/\partial Z^\alpha$ and hence passing through the origin of $\mathbb{T}$. Another way of viewing this cross-section is as a map q from $\tilde{\$}^*$ to $\mathbb{T}$ such that the composition poq is the identity on $\tilde{\$}^*$.

496 On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

This is one method of "splitting" the exact sequence; there is a corresponding dual method which is simply a map j from $\mathbb{T}$ to $\mathbb{S}$ with the property that the composition $j \circ i$ is the identity on $\mathbb{S}$. In the flat case, the maps j and q contain the same information as each other, so they can each equivalently be used to represent a (complex) space-time point. However, when $\mathbb{T}$ is suitably (mutilated and) deformed, the information will be different. The map q now gives the standard representation of a point in the right-flat non-linear graviton construction; and some corresponding "deformed" version of the map j ought to give a different difinition of a "point," the space of these "points" providing the required left-flat complex space-time.

There are many difficulties in making such an idea work in detail, however. It seems that we need some non-compact version of Kodaira's theorem [14,15] in order to be sure that the family of deformed j-maps is actually 4-dimensional. We need some suitable (and yet unknown) regularity conditions in the neighborhood of the image of i. Indeed, it appears that this neighborhood contains the entire information of the self-dual curvature. We need, also, to be able to see why this neighborhood, if only infinitesimally deformed, is characterized by a 1st cohomology group of $\mathcal{O}(-6)$ twistor functions. Work is in progress on these problems.

The idea, in order to proceed further, would be that with one and the same deformed twistor space we might have three alternative definitions of "space-time point," the above two and a third which would be some kind of symmetrical combination of the other two, providing a general solution of Einstein's vacuum equations! Assuming this solution to be asymptotically flat, the q-definition would provide its $\tilde{\mathrm{H}}$-space and the p-definition its H-space [16]. At the moment, however, this is all purely speculative.

References

1	R. Penrose,	"Twistor Theory, Its Aims and Achievements" in Quantum Gravity, eds: C. J. Isham, R. Penrose, D. W. Sciama, Oxford University Press, 1975.
2	R. Penrose,	"Twistors and Particles" in Quantum Theory and the Structures of Time and Space, eds: L. Castell, M. Drieschner, C. F. von Weiszacker, C. Hanser, Munich 1975.
3	J. Morrow and K. Kodaira,	Complex Manifolds; Holt, Rinehart and Winston, N. Y., 1971.
4	R. C. Gunning and H. Rossi,	Analytic Functions of Several Complex Variables, Prentice-Hall, 1965.
5	R. O. Wells, Jr.,	Differential Analysis on Complex Manifolds, Prentice-Hall, 1973.
6	G. A. J. Sparling and R. Penrose,	"The Twistor Quadrille" in Twistor Newsletter 1(4 March 1976).
7	N. M. J. Woodhouse,	"Twistor Cohomology Without Sheaves," in Twistor Newsletter 2(10 June 1976).
8	R. Penrose	Gen. Rel. Grav. 7, 31, 171(1976).
9	R. S. Ward,	"The Twisted Photon" in Twistor Newsletter 1(4 March 1976).
10	R. S. Ward,	"Zero-rest-mass Fields from Twistor Functions" in Twistor Newsletter 1 (4 March 1976).
11	R. Penrose,	Proc. Roy. Soc. A284, 159(1965).
12	L. P. Hughston,	"A Generalized Right-Handed Photon Construction" in Twistor Newsletter 3 (7 December 1976).
13	R. Penrose and M. A. H. MacCallum,	Phys. Reports 6C, 241(1973).
14	K. Kodaira,	Am. J. Math. 85, 79(1963).
15	K. Kodaira,	Ann. Math. 75, 146(1962).
16	R. Hansen, E. T. Newman, R. Penrose and K. P. Tod,	Proc. Roy. Soc. (to appear).

 On the twistor descriptions of massless fields. *Complex manifold techniques in theoretical physics (Proc. Workshop, Lawrence, Kan., 1978).* Eds D. E. Lerner and P. D. Sommers, *Res. Notes in Math.*, Vol. 32, Pitman, Boston, 1979, pp. 55–91

17 L. P. Hughston, "The Twistor Cohomology of Local Hertz Potentials" in Twistor Newsletter 4 (1 April 1977).

18 D. Lerner, "The 'Inverse Twistor Function' for Positive Frequency Fields" in Twistor Newsletter 5(11 July 1977).

R PENROSE
Mathematical Institute
24-29 St. Giles
Oxford, England

110
COMBINATORIAL QUANTUM THEORY AND QUANTIZED DIRECTIONS

Advances in twistor theory. Ed. L. P. Hughston, Pitman, 1979

The article that formed the appendix to my Adams Prize Essay (Chapter 28)—and which was originally the main part of a Fellowship Report to St John's College Cambridge while I was a Research Fellow there during the years 1957–1959—is here presented as a separate self-contained account. See also Chapters 31, 46, and 56.

500 Combinatorial quantum theory and quantized directions. *Advances in twistor theory.* Ed. L. P. Hughston, Pitman, 1979

§5.14 COMBINATORIAL QUANTUM THEORY AND QUANTIZED DIRECTIONS *by R. Penrose*

According to conventional quantum theory, angular momentum can take only integral values (measured in units of $1/2\ \hbar$) and the (probabilistic) rules for combining angular moments are of a combinatorial nature. Also, according to quantum theory, a system with zero total angular momentum must be spherically symmetrical. A system must, thus, involve a relatively large angular momentum in order to determine a well-defined direction in space, and we may picture the axis of the angular momentum as giving such a direction. In the limit of large angular momenta, we may therefore expect that the quantum rules for angular momentum will determine the geometry of directions in space. We may imagine these directions to be determined by a number of spinning bodies. The angles between their axes can then be defined in terms of the probabilities that their total angular momenta (i.e. their "spins") will be increased or decreased when, say, an electron is thrown from one body to another. In this way the geometry of directions may be built up, and the problem is then to see whether the geometry so obtained agrees with what we know of the geometry of space and time.

Although there is a standard procedure for the treatment of such a problem (in the non-relativistic case: namely the use of 3-j and 6-j symbols, etc.), it is convenient to make use of an alternative (but equivalent) formalism which can be described very briefly as follows. Consider, first, an n-dimensional Kronecker delta δ_{ab}. Then a set of "isotropic" Cartesian tensors can be built up from this one symbol and scalars by means of the operations of addition, outer multiplication, transposition of indices, and contraction. The only identities satisfied by δ_{ab} which hold independently of n are $\delta_{ab} = \delta_{ba}$ and $\delta_{ab}\,\delta_{bc} = \delta_{ac}$ (summation convention assumed). Also, we have $\delta_{aa} = n\ (> 0)$ and a certain identity, which depends on n,

holds. Now consider an abstract structure which behaves formally exactly like the above set of tensors except that the formal equation $\delta_{aa} = -2$ is imposed, together with the one identity: $\delta_{ab}\,\delta_{cd} + \delta_{ac}\,\delta_{db} + \delta_{ad}\,\delta_{bc} = 0$. The "scalars" of this abstract structure will be taken to be the *rational* numbers. The tensor-like quantities which can be constructed in this way I call *binors*. It can then be shown that a necessary and sufficient condition for a binor $A_{ab\ldots d}$ to vanish is that its *norm* $\|A \ldots\| = A_{ab..d}A^{ab..d}$ (a rational number) should vanish.

To obtain the physical interpretation of the binors, I envisage the following type of situation. Imagine the universe to be represented as a

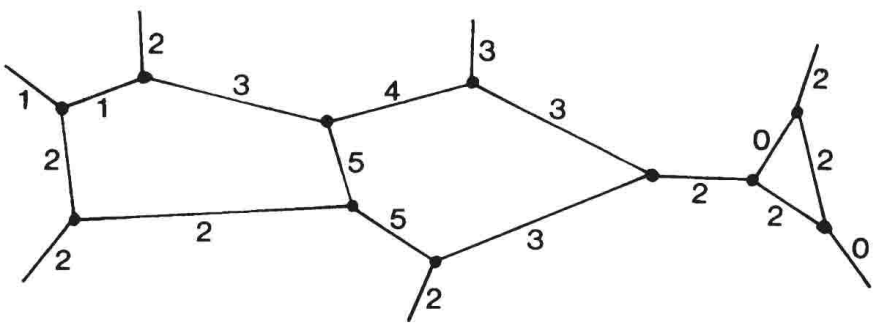


Figure A

network of segments (see Fig. A) where, for simplicity, it will be assumed that each internal segment connects two vertices and each vertex joins just three segment ends. Associated with each segment is a non-negative integer. Each segment is to be thought of as representing the world line of a particle, nucleus, atom, etc. Or, generally, some structure which may be momentarily considered as stationary and isolated from the rest of the

universe. The network thus gives a kind of combinatorial space-time picture of the universe. The integer associated with each segment is twice the total intrinsic angular momentum quantum number of the particle (or structure). Hence, for a pi-meson or ground state helium atom this integer is zero; for an electron or proton it is one; for a deuteron it is two; and so on. In order for the proposed calculus to give an accurate description of a part of the universe, it must be supposed that it is possible to neglect effects due to relative velocities of the particles (or structures). In this sense the theory is a non-relativistic one.

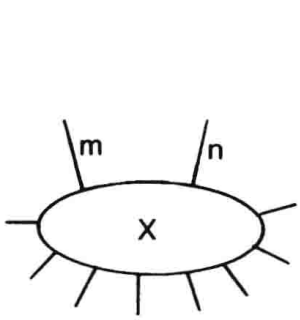
Corresponding to each possible such (open) network is associated a binor which is a contracted product of binors described as follows. Each segment numbered 0 is represented by the scalar 1. Each segment numbered 1 is represented by a δ_{ab}. Segments marked 2 by $\delta_{ac,bd} = 1/2!\ (\delta_{ab}\,\delta_{cd} - \delta_{ad}\,\delta_{cb})$, marked 3 by

$$\delta_{abc,\,def} = 1/3! \begin{vmatrix} \delta_{ad} & \delta_{ae} & \delta_{af} \\ \delta_{bd} & \delta_{be} & \delta_{bf} \\ \delta_{cd} & \delta_{ce} & \delta_{cf} \end{vmatrix}$$

and so on. The first group of indices of $\delta_{ab..d,ef..h}$ is to be associated with one end of the segment and the second group with the other end. (This is symmetrical since $\delta_{ab} = \delta_{ba}$ implies $\delta_{ab..d,ef..h} = \delta_{ef..h,ab..d}$.) At each vertex the three relevant groups of indices involved must all be paired off and contracted so that none of these indices remains uncontracted, and no two belonging to the same $\delta...,...$ are contracted together. This implies that the sum of the three integers involved is even, and is at least twice the greatest of them. Then, the result of these contractions will be unique up to sign. The free indices of the resultant binor are then just those

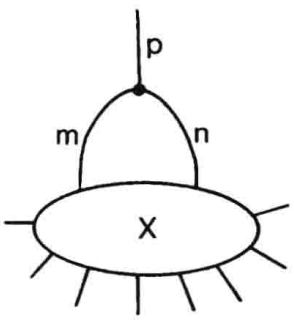
corresponding to the free ends of the network.

Consider, now, a situation in which a portion of the universe is represented by a known network X (see Fig. B) having, among its free ends, one numbered m and one numbered n. Suppose the particles or structures represented by these two ends combine together to form a new structure. (See Fig. C). We wish to know, for any allowable p, what is the probability



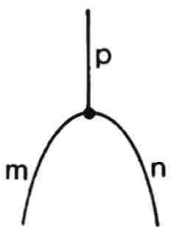
B....

Figure B



C....

Figure C



D....

Figure D

E....

Figure E

that the angular momentum number of this new structure be p. Fig. D is the network representing the combining of these two structures to form the third, but ignoring the rest of the universe; Fig. E is the "network" for the final structure alone. Let B..., C..., D..., E..., be the binors representing the networks of Figs. B, C, D and E, respectively. Then,

$$\textit{required probability} = \frac{\|\,C...\,\|}{\|\,B...\,\|} \times \frac{\|\,E...\,\|}{\|\,D...\,\|}$$

(in fact $\| E... \| = (-1)^p(p+1)$ and $\| D... \| = (-1)^{(m+n+p)/2}(\frac{m+n-p}{2})!$ $(\frac{n+p-m}{2})!$ $(\frac{p+m-n}{2})!$ $(\frac{m+n+p+2}{2})!/m!n!p!)$. As a particular case of this we can deduce the result that the binor corresponding to a network vanishes if and only if the situation is "forbidden" according to the rules of quantum theory.

Consider now the situation of Fig. F involving two bodies with *large* angular momenta M and N. We might define the *angle* between the axes of the

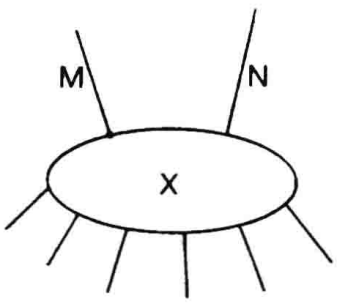


Figure F

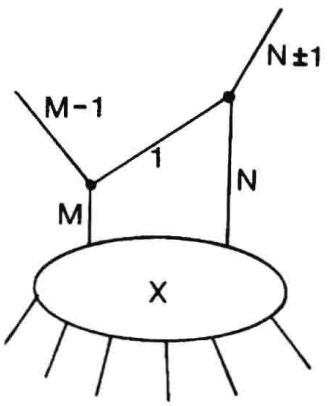


Figure G

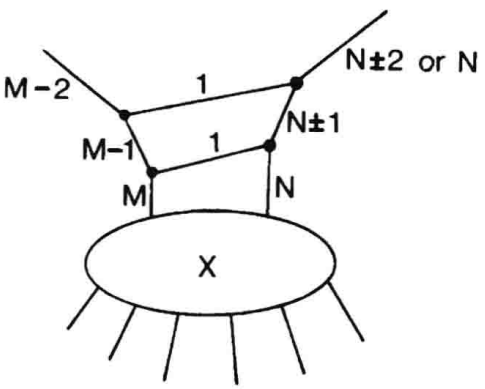


Figure H

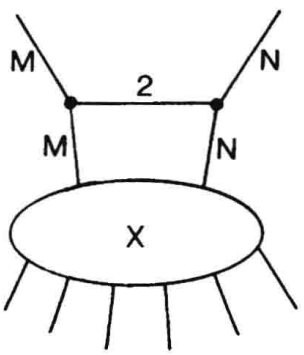


Figure I

bodies in terms of the relative probability (defined in terms of an ensemble of systems with the same Fig. F network) of occurrence of N+1 and N-1, respectively, in the "experiment" given by Fig. G. However, part of this probability may be due to *ignorance* of the relationship between the bodies. (This may manifest itself in the absence of sufficient connecting links in the known network X.) To eliminate the possibility that part of the probability be due to ignorance we envisage a repetition of the "experiment" as given in Fig. H. If the probability given in the second experiment is essentially unaffected by the result of the first experiment then we may say that the angle θ between the axes of the bodies is well-defined and is determined by this probability. It then turns out that the binor of Fig. I is essentially 1/2 cos θ times the binor of Fig. F and from this fact, and certain binor identities, it is possible to show that the angles obtained in this way satisfy the same laws as do angles in a three-dimensional Euclidean space.

This is very satisfactory and is perhaps a little surprising in at least two respects. In the first instance, since no complex numbers were used in the binor calculus it is somewhat remarkable that a full array of directions in three-dimensional space has been built up, rather than in, say, just a two-dimensional subspace, since in ordinary quantum theory, in order to build up all the wave functions for all the possible spin directions for an electron, *complex* linear combinations must be used. Secondly, according to standard quantum theory, the wave function of a system of high angular momentum will *not* normally determine a well-defined axis in space contrary to what has apparently been assumed here.

The answer to both these points seems to lie in the fact that the "directions" that emerge in the theory described here are things which are

defined by the systems in relation to one another and they will not generally agree with the directions in a *previously* given (and unnecessary!) background space. The space that is obtained here is to be thought of (indeed *must* be thought of) as being the one *determined* by the systems themselves.

It is to be hoped that some modification to the above scheme might enable the effects due to relative velocities of systems to be taken into account so that perhaps a four-dimensional space-time might be constructed. (Time is absent from the above theory even to the extent that the time ordering of events is irrelevant!) Two additional features would have to be involved. The first is that the relativistic addition of angular momenta includes the possibility of multiple pair creation and many of the additional complications implied by relativistic field theory. The second is that relative velocity implies the possibility of a mixing of spin with orbital angular momentum so that the idea of "distance" between the world lines or particles is involved. Particularly because of this second feature, the fully relativistic theory would seem to be of a different order of difficulty from the one treated above.

REFERENCES

Penrose, R. 1971a. In *Combinatorial Mathematics and its Applications*, ed. D.J.A. Welsh. Academic Press.

Penrose, R. 1971b. In *Quantum Theory and Beyond*, ed. T. Bastin. Cambridge University Press.

111
PENTAPLEXITY: A CLASS OF NONPERIODIC TILINGS OF THE PLANE

Mathematical Intelligencer **2** (1979/80), no.1, 32–7. Reprinted from *Eureka* **39** (1978), 16–22; also reprinted in *Geometrical Combinatorics* (1984), 55–65, *Res. Notes in Math.*, 114, Pitman, Boston, 1984

This article, like Chapters 1 and 2, appeared first in the Cambridge University student's mathematical magazine *Eureka* in 1978. Figures 6–8 of this article repeat Figures 4–6 of Chapter 64, but it is then shown how the aperiodic set of six tiles of Chapter 64 can be reduced to aperiodic sets with only two tiles, the so-called 'Penrose tiles', which were first published in a 1977 *Scientific American* article by Martin Gardner. See also Chapters 168 and 187.

 Pentaplexity: a class of nonperiodic tilings of the plane. *Mathematical Intelligencer* **2** (1979/80), no.1, 32–7. Reprinted from *Eureka* **39** (1978), 16–22; also reprinted in *Geometrical Combinatorics* (1984), 55–65, *Res. Notes in Math.*, 114, Pitman, Boston, 1984

Pentaplexity*

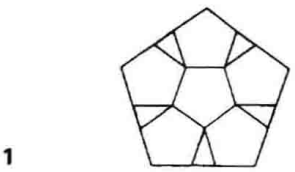
A Class of Non-Periodic Tilings of the Plane

R. Penrose

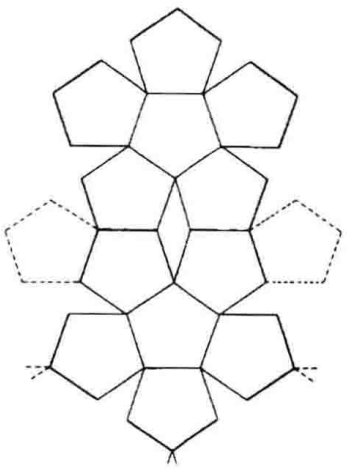
Some readers may be acquainted with an article by Martin Gardner in the January 1977 issue of Scientific American. In this he described a pair of plane shapes that I had found, called "kites" and "darts", which, when matched according to certain simple rules, could tile the entire plane, but only in a non-periodic way. The tilings have a number of remarkable properties, some of which were described in Gardner's article. I shall give here a brief account explaining how these tiles came about and indicating some of their properties.

The starting point was the observation that a regular pentagon can be subdivided into six smaller ones, leaving only five slim triangular gaps. (See Fig. 1; this is familiar as part of the usual "net" which folds into a regular dodecahedron.) Imagine, now, that this process is repeated a large number of times, where at each stage the pentagons of the figure are subdivided according to the scheme of Fig. 1. There will then be gaps appearing of varying shapes and we wish to see how best to fill these. At the second stage of subdivision, diamond-shaped gaps appear between the pentagons (Fig. 2). At the third, these diamonds grow "spikes", but it is possible to find room, within each such "spiky diamond", for another pentagon, so that the gap separates into a star (pentagram) and a "paper boat" (or jester's cap?) (Fig. 3). At the next stage, the star and the boat also grow "spikes", and, likewise, we can find room for new pentagons within them, the remaining gaps being new stars and boats (as before). These subdivisions are shown in Fig. 4.

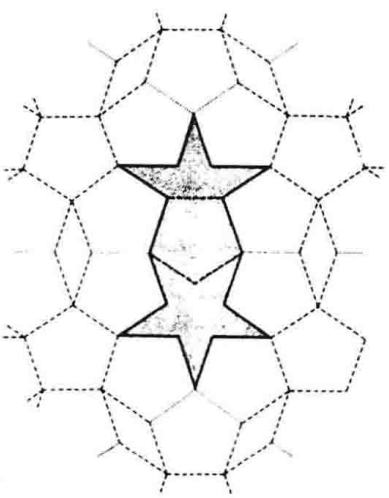
Since no new shapes are now introduced at subsequent stages, we can envisage this subdivision process proceeding indefinitely. At each stage, the scale of the shapes can be expanded outwards so that the new pentagons that arise become the same size as those at the previous stage. As things stand, however, this procedure allows an ambiguity that we would like to remove. The subdivision of a "spiky diamond" can be achieved in two ways, since there are two alternative positions for the pentagon. Let us insist on just *one* of these, the rule being that given in Fig. 5. (When we examine the pattern of surrounding pentagons we necessarily find that they are arranged in the type of configuration shown in Fig. 5.) It may be mentioned that had the opposite rule been adopted for subdividing a "spiky diamond", then a contradiction would



1

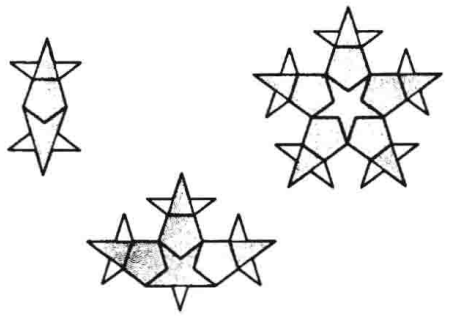


2

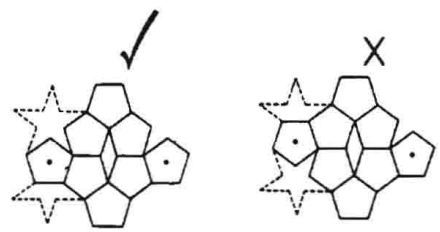


3

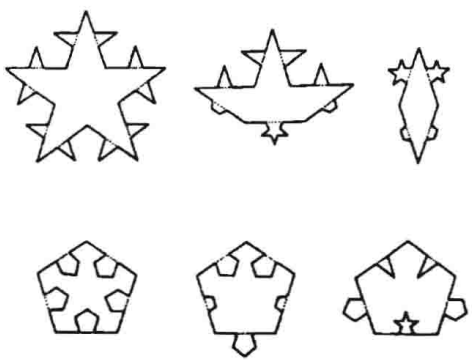
* The INTELLIGENCER thanks the Archimedeans of Cambridge University for permission to reprint *Pentaplexity* which first appeared in *Eureka* No. 39.



4



5



6

appear at the next stage of subdivision, but this never happens with the rule of Fig. 5.

This procedure, when continued to the limit, leads to a tiling of the entire plane with pentagons, diamonds, boats and stars. But there are many "incorrect" tilings with these same shapes, being not constructed according to the above prescription. In fact, "correctness" can be *forced* by adopting suitable matching rules. The clearest way to depict these rules is to modify the shapes to make a kind of infinite jigsaw puzzle, where a suggested such modification is given in Fig. 6. It is not too hard to show that any tiling with these six shapes is forced to have a hierarchical structure of the type just described.

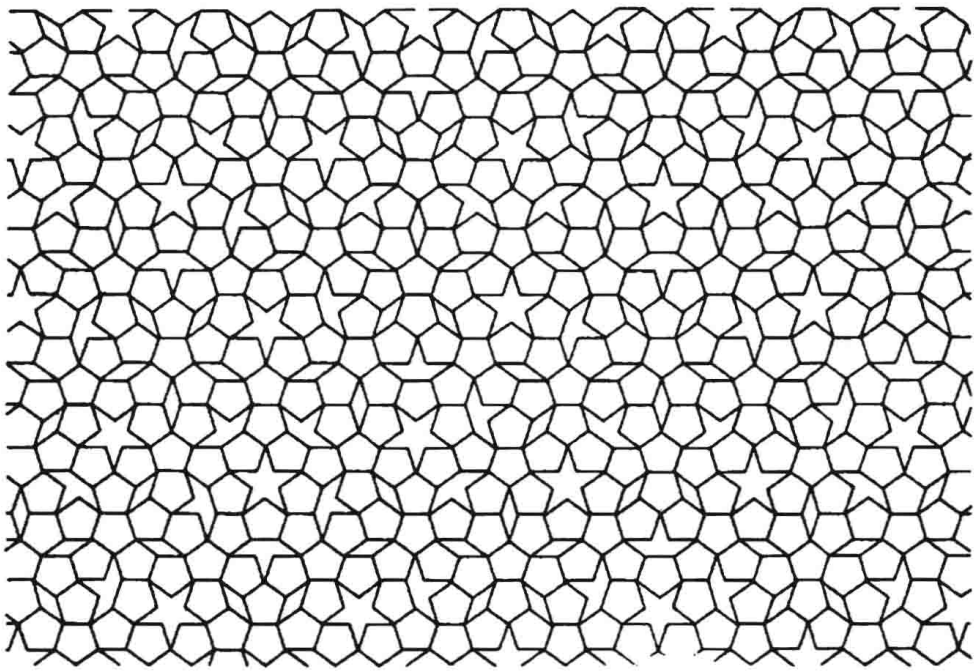
Furthermore, the forced hierarchical nature of this pattern implies that the tiling has a number of very remarkable properties. In the first place, it is necessarily *non-periodic* (i.e. without any period parallelogram). More about this later. Secondly, though the completed pattern is not uniquely determined – for there are $2^{\aleph_0}$ different arrangements – these different arrangements are, in a certain "finite" sense, all indistinguishable from one another! Thus, no matter how large a finite portion is selected in one such pattern, this finite portion will appear somewhere in *every* other completed pattern (infinitely many times, in fact). Thirdly, there are many unexpected and aesthetically pleasing features that these patterns exhibit (see Fig. 7). For example, there are many regular decagons appearing, which tend to overlap in places. Each decagon is surrounded by a ring of twelve pentagons, and there are larger rings of various kinds also. Note that every straight line segment of the pattern extends outwards indefinitely, to contain an infinite number of other line segments of the figure. The hierarchical arrangement of Fig. 7 is brought out in Fig. 8.

After I had found this set of six tiles that forces non-periodicity, it was pointed out to me (by Simon Kochen) that Raphael Robinson had, a number of years earlier, also found a (quite different) set of six tiles that forces non-periodicity. But it occurred to me that with my tiles one could do better. If, for example, the third "pentagon" shape is eliminated by being joined at two places to the "diamond" and at one place to the bottom of the "boat", then a set of *five* tiles is obtained that forces non-periodicity. It was not hard to reduce this number still further to four. And then, with a little slicing and rejoining, to *two*!

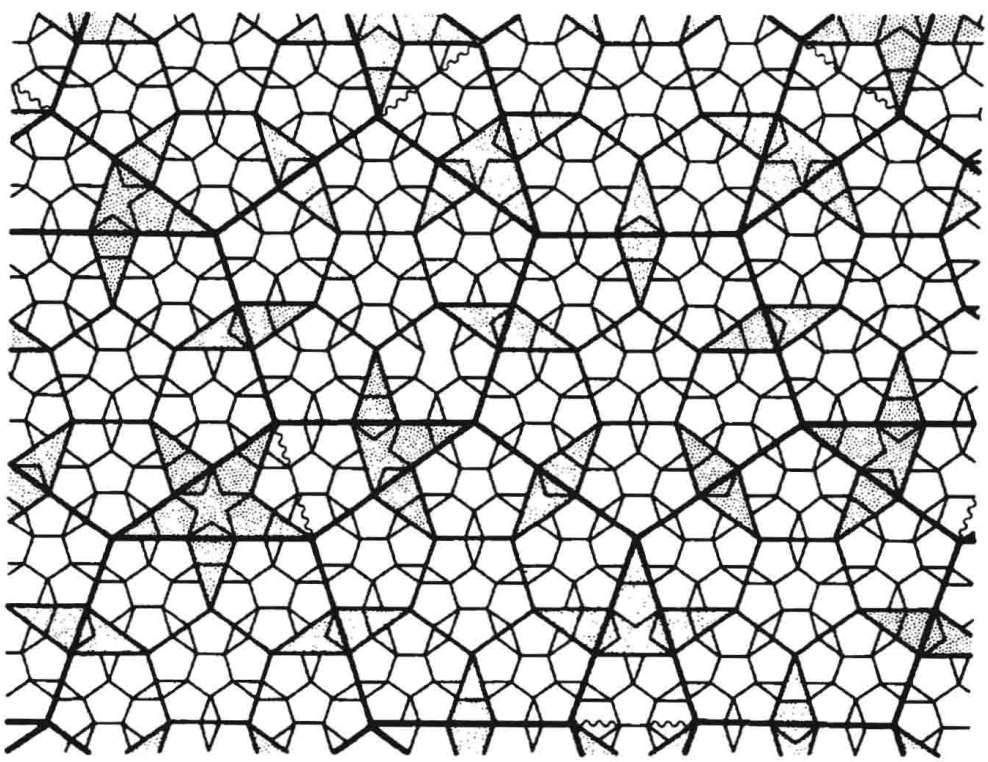
The two tiles so obtained are the "kites" and "darts" mentioned at the beginning*. The precise shapes are illustrated in Fig. 9. The matching rules are also shown, where vertices of the same colour must be placed against one another. There are many alternative ways to colour or shade these tiles to force the correct arrangements. One way which brings out the relation to the pentagon-diamond-boat-star tilings is shown in Fig. 10. A patch of

* These names were suggested by John Conway.

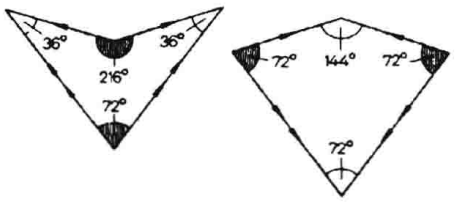
510 Pentaplexity: a class of nonperiodic tilings of the plane. *Mathematical Intelligencer* **2** (1979/80), no.1, 32–7. Reprinted from *Eureka* **39** (1978), 16–22; also reprinted in *Geometrical Combinatorics* (1984), 55–65, *Res. Notes in Math.*, 114, Pitman, Boston, 1984



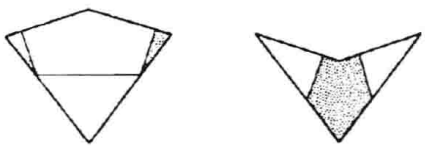
7



8



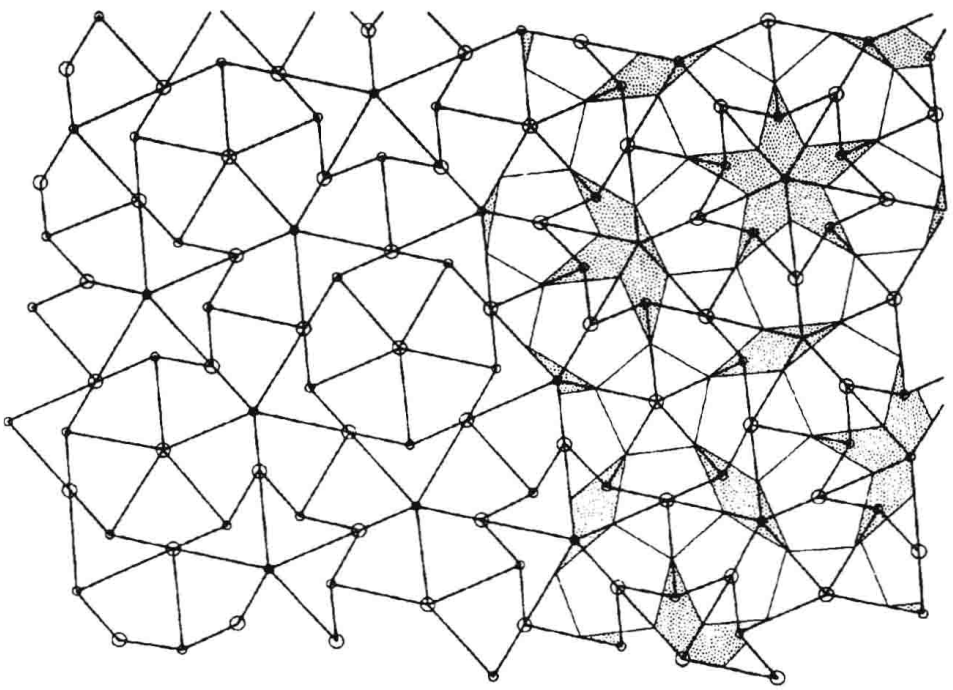
9



10

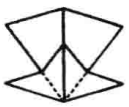
assembled tiles (partly shaded in this way) is shown in Fig. 11.

The hierarchical nature of the kite-dart tilings can be seen directly, and is illustrated in Fig. 12. Take any such tiling and bisect each dart symmetrically with a straight line segment. The resulting half-darts and kites can then be collected together to make darts and kites on a slightly larger scale: two half-darts and one kite make a large dart; two half-darts and two kites make a large kite. It is not hard to convince oneself that every correctly matched kite-dart tiling is assembled in this way. This "inflation" property also serves to prove non-periodicity. For suppose there were a period parallelogram. The corresponding inflated kites and darts would also have to have the same period parallelogram. Repeat the inflation process many times, until the size of the resulting inflated kites and darts is greater than that of the supposed period parallelogram. This gives a contradiction.

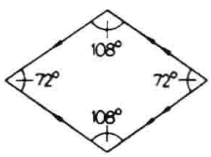


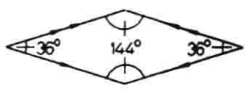
11

 Pentaplexity: a class of nonperiodic tilings of the plane. *Mathematical Intelligencer* **2** (1979/80), no.1, 32–7. Reprinted from *Eureka* **39** (1978), 16–22; also reprinted in *Geometrical Combinatorics* (1984), 55–65, *Res. Notes in Math.*, 114, Pitman, Boston, 1984

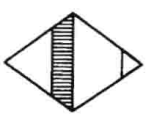


12

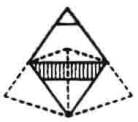




13



14



15

The contradiction with periodicity shows up in another striking way. Consider a very large area containing d darts and k kites, which is obtained referring to the inflation process a large number of times. The larger the area, the closer the ratio $x = k/d$ of kites to darts will be to satisfying the recurrence relation $x = (1 + 2x)/(1 + x)$ (since, on inflation, a dart and two kites make a larger kite, while a dart and a kite make a larger dart). This gives, in the limit of an infinitely large pattern, $x = \frac{1}{2}(1 + \sqrt{5}) = \tau$, the golden ratio! Thus we get an *irrational* relative density* of kites to darts – which is impossible for a periodic tiling.

There is another pair of quadrilaterals which, with suitable matching rules, tiles the plane only non-periodically. This is a pair of rhombuses shown in Fig. 13. In Fig. 14 a suitable shading is suggested where similarly shaded edges are to be matched against each other. In Fig. 15, the hierarchical relation to the kites and darts is illustrated. The rhombuses appear mid-way between one kite-dart level and the next inflated kite-dart level.

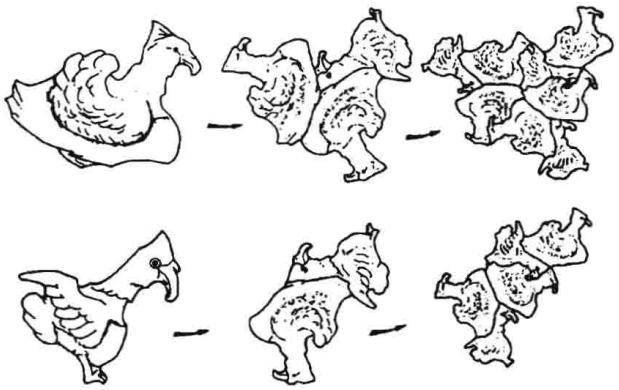
Many different jigsaw puzzle versions of the kite-dart pair or the rhombus pair can evidently be given. One suggestion for modified kites and darts, in the shape of two birds, is illustrated in Fig. 16. The inflation process (in reverse) is illustrated in Fig. 17.

Other modifications are also possible, such as alternative matching rules, suggested by Robert Ammann (see Fig. 18) which force half the tiles to be inverted.

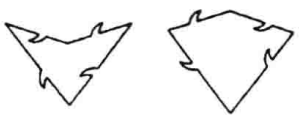
Many intriguing features of the tilings have not been mentioned here, such as the pentagonally-symmetric rings that the stripes of Fig. 14 produce, Conway's classification of "holes" in kite-dart patterns (i.e. regions surrounded by "legal" tilings but which cannot themselves be legally

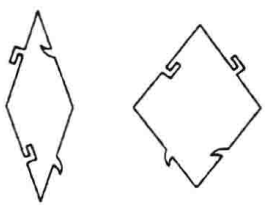
* This is the *numerical* density. The kite has τ times the area of the dart, so the total area covered by kites is τ^2 $(= 1 + \tau)$ times that covered by darts.

16

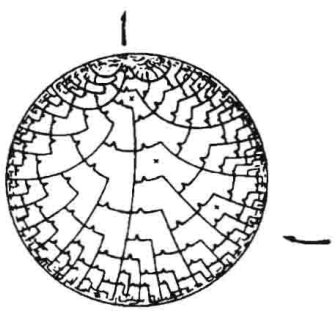


17





18



19

filled), Ammann's three-dimensional version of the rhombuses (four solids that apparently fill space only non-periodically), Ammann's and Conway's analysis of "empires" (the infinite system of partly disconnected tiles whose positions are forced by a given set of tiles).

It is not known whether there is a *single* shape that can tile the Euclidean plane only non-periodically. For the hyperbolic (Lobachevski) plane a single shape can be provided which, in a certain sense, tiles only non-periodically (see Fig. 19) – but in another sense a periodicity (in one direction only) can occur. (This remark is partly based on a suggestion of John Moussouris).

References

Gardner, M., *Scientific American* January 1977, pp. 110–121.
Penrose, R., *Bull. Inst. Maths. & its Applns.* 10, No. 7/8 (1974) pp. 266–271.

R. Penrose
Mathematical Institute
University of Oxford
28, St. Giles
Oxford, U. K.

112
RECENT ADVANCES IN GLOBAL GENERAL RELATIVITY: A BRIEF SURVEY

Einstein Symposium, Berlin (1979), pp. 36–45; *Lecture Notes in Phys.*, **100**, Springer, Berlin-New York, 1980

Numerous strict conformal diagrams appear here, depicting well-known space-times.

516 Recent advances in global general relativity: a brief survey. *Einstein Symposium, Berlin* (1979), pp. 36–45. *Lecture Notes in Phys.*, **100**, Springer, Berlin-New York, 1980

RECENT ADVANCES IN GLOBAL GENERAL RELATIVITY: A BRIEF SURVEY

Roger Penrose, Mathematical Institute, Oxford, U.K.

Theoretical research into Einstein's general relativity has made some very significant advances over the past 20 years or so. During this period, attention has been concentrated almost entireley on Einstein's original 1915 theory rather than on the various later attempts to generalize it to some kind of unified field theory. This research has revealed a number of desirable and remarkable properties, both physical and mathematical, that the original theory possesses. (It is hard to see that many of these could survive if the theory were to be generalized, unless new insights, as deep as those that led Einstein to his 1915 theory, may eventually be invoked in making such a generalization.)

Many of these recent advances have been into the global rather than the local structure of the theory. Early work into general relativity had, to a large degree, been concerned with finding exact solutions to the field equations - always a difficult task owing to the complicated non-linear nature of these equations and to the fact that establishing equivalence between two such solutions depends upon making general co-ordinate transformations. In the study of global questions, however, one is not concerned with particular exact solutions but with general qualitative properties. The detailed solution of Einstein's equation becomes of secondary importance and attention is concentrated, instead, on geometrical and topological properties of space-times.

One of the topics that has received a great deal of attention in recent years is the study of black holes. Here, exact solutions, notably the Schwarzschild solution and its generalization to that of Kerr, *do* also play important roles, first as models indicative of the kind of behaviour to be expected in more general situations, and second, remarkably, as the final states into which *all* black holes finally settle down.

The familiar Schwarzschild solution, as extended to within the Schwarzschild radius $r = 2m$ by means of, say, Eddington-Finkelstein coordinates [1,2], provides the basic (spherically symmetrical) model of the final collapse of a massive star to form a black hole (see Fig. 1). This model was put forward by Oppenheimer and Snyder [3] in 1939, but it had not been clear to what extent its two salient features, namely the horizon at $r = 2m$ and the curvature singularity at $r = 0$, were stable properties under perturbations of the model or mere artefacts of the

very special assumption of exact spherical symmetry.

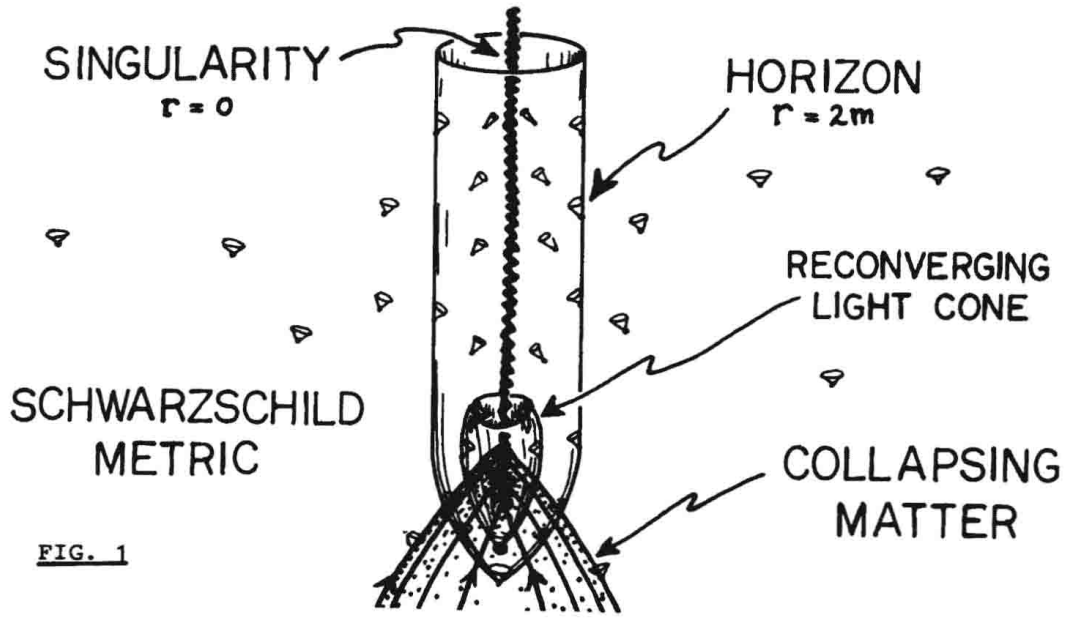


FIG. 1

In fact, with regard to the singularity it has been possible to show [4,5,6] using global techniques of differential topology, that the presence of a space-time singularity is, indeed, a stable property. For this one requires a suitable criterion signifying that, in some suitable sense, collapse past a point of no return has actually taken place. The existence of a "trapped surface" is one such criterion [4], and of a point whose future light cone eventually begins to reconverge in every direction [5], is another. These criteria can also be used in the reverse direction in time in order to establish that the big-bang singularity, also, is not merely a feature of the high symmetry that is assumed for the standard cosmological models, but is a rigorous consequence of Einstein's thoery [7,5,6] as applied to (seemingly) very convincing astrophysical data (such as the nature of the 3^0K background radiation). Such singularity theorems can also be used in other contexts and are very broadly [6] applicable. But they are somewhat negative in the sense that they give very little hint as to the detailed nature of the predicted singularities. In recent years there have been some improvements in this (and other) respects [8], over the original results.

The status of the existence of a horizon in asymmetric gravitational collaps is not so clear however. The question is whether the singularities that arise in collapse can ever be "naked", *i.e.* visible to an observer at large distances. If there is some general principle preventing this (according to the laws of classical general relativity), at

least in cases of "generic" collapse, then we say that *cosmic censorship* holds true [9], and the forming of horizons about singularities is a necessary consequence of the theory. But at present, despite some encouraging partial evidence to support it, cosmic censorship remains an unproved conjecture [8] probably the most important unsolved problem of classical general relativity theory.

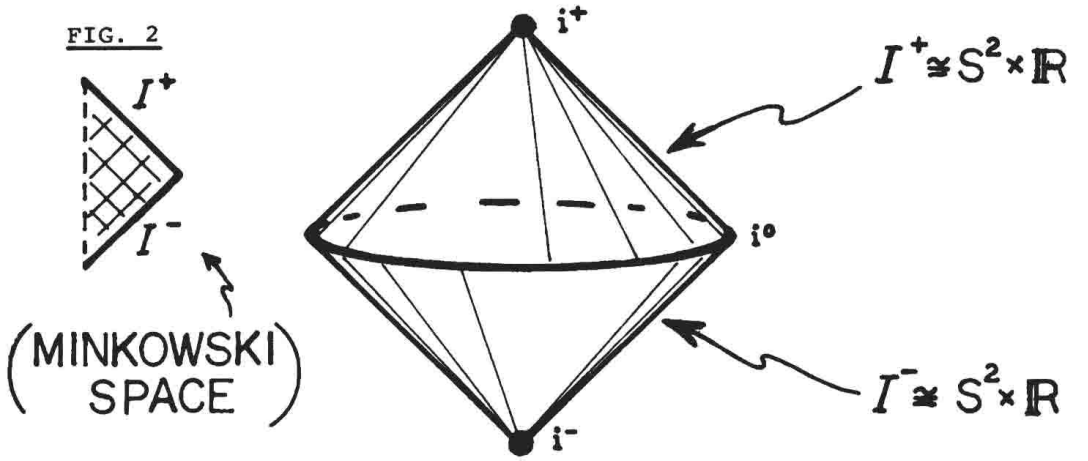
The general argument that irretrievable gravitational collapse always results in a black hole rather than in a naked singularity does, strictly speaking, require the cosmic censorship assumption. However, most astrophysicists seem quite prepared to accept the former possibility as the "lesser of two evils" - and what theoretical evidence that there is seems, so far, to support them. Having accepted that black holes can exist in very general situations of gravitational collapse, it may be further presumed that, if left undisturbed on its own, a black hole will settle down into a stationary configuration. Granting this, one can then appeal to a succession of remarkable theorems [10-13] which, in effect, establish the precise form of the resulting space-time metric as that known as the *Kerr metric* [14,6], depending upon only two parameters a and m , where m refers to the mass and am , to the angular momentum (a = 0 giving the Schwarzschild metric). The gravitational field of the initial collapsing body would require many more parameters (infinitely many, in principle,) for its detailed description, but almost all this information gets washed out in the final black-hole state, the quadrupole and higher multipole moments being radiated away in gravitational radiation emitted as the configuration finally settles down to that of Kerr.

An unexpected bonus, from the computational point of view, is that the geodesic equations can be reduced simply to quadratures in the Kerr metric [15], owing to the presence of an extra constant of the motion for particle orbits (Carter's constant), and this helps immensely in the analysis of the metric. A further constant for null orbits enables the rotation of light polarization to be integrated explicitly [16]. Furthermore, the wave equation [15] and its generalization for all the various spins [17,18], and also the Dirac equation [19], all separate in the Kerr background, so that the solution of these equations becomes a comparatively straightforward matter. All this seems a remarkable gift from Einstein's 1915 theory!

When gravitational waves are emitted by an isolated system, some energy is lost by the system. Indeed, it seems to have been this demonstration, by Bondi [20] and Sachs [21] in around 1960, that gravitational waves

carry positive mass-energy, that largely initiated the renewed interest into general relativity of more recent years. Since gravitational energy cannot be localized, the study of such questions requires some form of asymptotic analysis. A convenient way of handling asymptotic questions in general relativity is the introduction of a conformal boundary to space-time, yielding a smooth conformal manifold-with-boundary in most of the cases of interest [22]. For a radiating isolated system this boundary consists of two pieces $\mathscr{I}^-$ and $\mathscr{I}^+$, each of which is a null hypersurface of topology $S^2 \times \mathbb{R}$ (see Fig. 2). The geometry and other data at $\mathscr{I}^-$ describe the incoming radiation fields and at $\mathscr{I}^+$, the outgoing radiation fields. In this way, the Bondi-Sachs energy-momentum law, and certain generalizations such as to incorporate other massless fields interacting with gravity, may be simply derived [23].

FIG. 2



Important problems remain to be solved, however, such as establishing, in an unambiguous way, the relation between the Bondi-Sachs energy-momentum and the more familiar energy-momentum integral [24,25] defined at spacelike infinity (though significant advances have been made recently [26]), the establishment of a "positive energy theorem" that would ensure that the total Bondi-Sachs energy is always non-negative (which would considerably strengthen the results already established [27] for the more familiar spacelike "positive energy conjecture"), and obtaining a proper understanding of angular-momentum and mass-centre in general relativity. The difficulty that arises in connection with this last problem occurs because the rotation elements of the Poincaré group cannot be uniquely identified as asymptotic symmetries of a general asymptotically flat space-time. The problem does not arise for the translation elements, which is why a unique energy-momentum concept does arise. But the concept of an asymptotic rotation is (retarded) time-dependent, so there

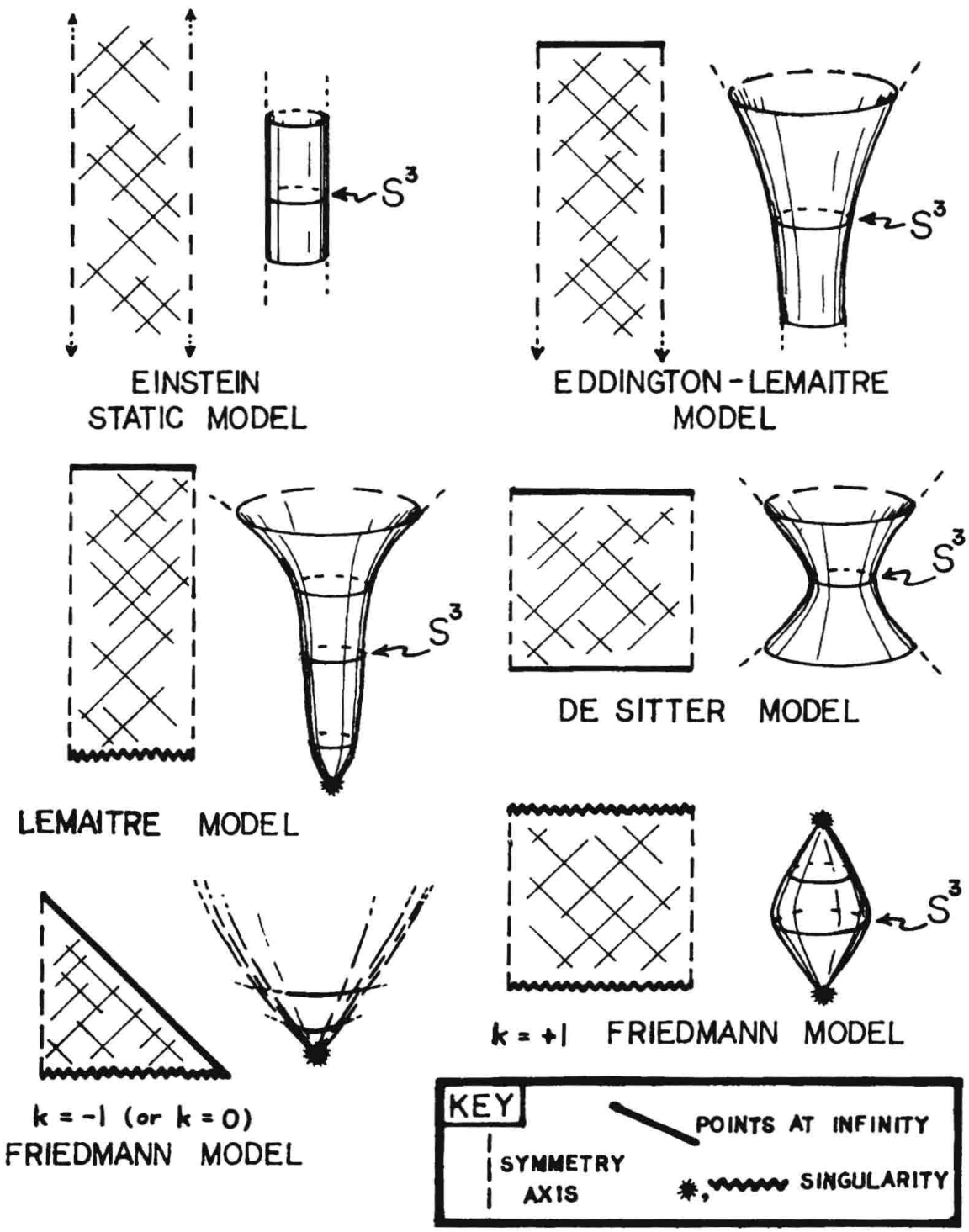
 Recent advances in global general relativity: a brief survey. *Einstein Symposium, Berlin* (1979), pp. 36–45. *Lecture Notes in Phys.*, **100**, Springer, Berlin-New York, 1980

seems to be no clear way of identifying the different angular momentum concepts that are defined for different (retarded) times [28,29].

The origin of all these difficulties is the fact that the group of asymptotic symmetries of an asymptotically flat space-time, defined as the appropriate conformal symmetries of $\mathscr{I}^+$ (or of $\mathscr{I}^-$), is not the Poincaré group, but an infinite-parameter group known as the Bondi-Metzner-Sachs (BMS) group [20,28], in which the Poincaré rotation elements cannot be uniquely singled out. This prompted Newman to initiate another approach to asymptotic "symmetries" involving a complexification of the asymptotic space-time parameters [30,31]. Although this has not, as yet, led to a clarification of angular momentum it has had a number of other interesting consequences. It provides a remarkable concept known as $\mathscr{H}$-space - which may be regarded as the complex "vacuum space-time" reconstructed solely from the self-dual part of the gravitational radiation field. This approach has some remarkable mathematical aspects to it, but its physical significance remains somewhat obscure. One of its important properties is that it provides a direct link with the theory of twistors [32] and points the way towards a general construction of non-linear fields in terms of deformations of twistor space (or of bundles over twistor space) [33,34].

Another instance of the value of introducing conformal boundaries to space-time is in the study of cosmological models. Here, the boundary can describe either a space-time singularity or infinity (Fig. 3). In the case of the big-bang singularity we obtain a conformal boundary which is a spacelike hypersurface. This provides a simple understanding of the *particle horizons* [35,36] that these models possess, whereby an observer in the early stages of the expansion can receive signals only from a limited, but steadily increasing, number of particles in the universe.

FIG. 3


S^3
EINSTEIN
STATIC MODEL
S^3
EDDINGTON - LEMAITRE
MODEL
S^3
LEMAITRE MODEL
S^3
DE SITTER MODEL
k = -1 (or k = 0)
FRIEDMANN MODEL
S^3
k = +1 FRIEDMANN MODEL
KEY
POINTS AT INFINITY
SYMMETRY
AXIS
SINGULARITY

 Recent advances in global general relativity: a brief survey. *Einstein Symposium, Berlin* (1979), pp. 36–45. *Lecture Notes in Phys.*, **100**, Springer, Berlin-New York, 1980

Several different methods [6,8] have been suggested for attaching boundaries to space-time. The one that I have been adopting here, the *causal boundary* [37], assigns the same future end-point to two timelike or null curves if and only if their (chronological) pasts coincide; and, accordingly, the same past end-point whenever their (chronological) futures coincide. Another method that had gained widespread acceptance is the b-boundary [38] (which provides only singular points and not points at infinity) and its conformal generalization [39] (which provides also points at infinity). This method has encountered certain difficulties [40], however, when applied to cosmological models, and it is hard to use in practice. The causal boundary has the advantage of simplicity and the fact that it relates directly to physically important concepts (*e.g.* horizons).

One of the implications of the cosmic censorships assumption is the area theorem for black holes [41,42]. This asserts that the surface area of the horizon is non-decreasing with time. Any disturbance to a black hole will serve to increase its surface area. For a Kerr hole, the formula for this area is

$$A = 8\pi m(m + (m^2 - a^2)^{1/2}) \ .$$

In the non-rotating case $a = 0$, so A is simply proportional to m^2. The mass of a non-rotating black hole thus grows whenever the hole is disturbed. This need not be so in the rotating case, however, and it is possible to extract some mass-energy from the hole at the expense of slowing down its rotation. This can be done either by suitable projecting particles into the hole [9,43] or by means of external fields ("superradiance") [44,45].

This irreversible behaviour of a black hole's horizon relates to the second law of thermodynamics. By means of thought experiments [46] and considerations of quantum field theory in a curved (black-hole) background [47] the Bekenstein-Hawking formula

$$S = \tfrac{1}{4} \frac{kAc^3}{\hbar G}$$

for the entropy of a black hole of surface area A has been derived (k = Boltzmann's constant, $\hbar$ = Planck's constant/2π, G = Newton's gravitational constant and c = velocity of light). This implies that a black hole should also have a non-zero temperature (inversely proportional to the mass, for a non-rotating hole) though for a black hole

arising in stellar collapse, this temperature would be utterly insignificant ($\sim 10^{-8}$ °K for a 10 $M_\odot$ hole). In principle, in an ever-expanding cosmological model, such a black hole would eventually radiate away all its mass because of this non-zero temperature (Hawking radiation) though this process would take over 10^{53} times the present age of the universe for a hole arising from stellar collapse [47]. On the other hand if very tiny holes (of mass $\sim 10^{15}$g and radius $\sim 10^{-13}$cm) were produced in a highly chaotic big bang, then their final explosive disappearance could have observational significance [48]. (It should be remarked that the area theorem of classical general relativity is being violated here, which is legitimate because the normal local positivity assumption for energy is violated for *quantum* fields.)

For reasons connected with the second law of thermodynamics and the nature of gravitational entropy, however, it is highly reasonable to infer that the big bang singularity was *not* of such a chaotic nature [42,49], so such small black holes would seem to be very improbable. (This is also consistent with direct observation [50].) A suitably strong version of the cosmic censorship assumption implies that singularities divide themselves into two quite distinct types: past and future. (This is essentially equivalent to the assumption that the universe is globally hyperbolic [42,49], which incidentally implies that the topology of the universe is unchanging with time [51].) The past-type singularities would apparently have to be uniform and unchaotic in nature while the future singularities would not be so constrained and would indeed be expected to be very chaotic and non-uniform.

No prior theoretical reason for such an asymmetrical behaviour of singularities has been put forward, however. It would seem that the answer to this puzzle (which would include an answer to the question of the origin of the second law of thermodynamics) must lie in the finding of a correct theory that unifies general relativity with quantum mechanics. Unfortunately, despite many heroic efforts, we seem as far away from such a theory as ever [52].

References

[1] A. S. Eddington, Nature 113 (1924) 192

[2] D. Finkelstein, Phys. Rev. 110 (1958) 965

[3] J. R. Oppenheimer and H. Snyder, Phys. Rev. 56 (1939) 455

[4] R. Penrose, Phys. Rev. Lett. 14 (1965) 57

[5] S. W. Hawking and R. Penrose, Proc. Roy. Soc. A314 (1970) 529

 Recent advances in global general relativity: a brief survey. *Einstein Symposium, Berlin* (1979), pp. 36–45. *Lecture Notes in Phys.*, **100**, Springer, Berlin-New York, 1980

[6] S. W. Hawking and G. F. R. Ellis, The Large Scale Structure of Space-time (Camb. Univ. Press 1973)

[7] S. W. Hawking and G. F. R. Ellis, Astrophysics J. 152 (1968) 25

[8] F. J. Tipler, C. J. S. Clarke and G. F. R. Ellis, Singularities and Horizons - a Review Article (to appear in G R G Einstein Centenary Volume 1979).

[9] R. Penrose, Revista del Nuovo Cimento, Serie I, 1 Numero Speciale (1969) 252

[10] W. Israel, Phys. Rev. 164 (1967) 1776

[11] B. Carter, Phys. Rev. Lett. 26 (1971) 331

[12] S. W. Hawking, in Black Holes: Les Astres Occlus, Les Houches Summer School 1972 (Eds. C. M. DeWitt and B. S. DeWitt, Gordon & Breach, New York 1973)

[13] D. C. Robinson, Phys. Rev. Lett. 34 (1975) 905

[14] R. P. Kerr, Phys. Rev. Lett. 11 (1963) 237

[15] B. Carter, Commun. Math. Phys. 10 (1968) 280

[16] M. Walker and R. Penrose, Commun. Math. Phys. 18 (1970) 265

[17] S. Teukolsky, Ann. New York Acad. Sci. 262 (1975) 275

[18] S. Chandrasekhar, Proc. Roy. Soc. A358 (1978) 421, 441

[19] S. Chandrasekhar, Proc. Roy. Soc. A349 (1976) 571

[20] H. Bondi, M. G. J. van der Burg and A. W. K. Metzner, Proc. Roy. Soc. A269 (1962) 21

[21] R. K. Sachs, Proc. Roy. Soc. A270 (1962) 103

[22] R. Penrose, Proc. Roy. Soc. A284 (1975) 159

[23] R. Penrose, in: Relativity Theory and Astrophysics, 1 (ed. J. Ehlers, Amer. Math. Soc. 1967)

[24] L. D. Landau and E. M. Liftschitz, The classical Theory of Fields (Pergamon Press, Oxford 1962)

[25] R. Arnowitt, S. Deser and C. W. Misner, in: Gravitation: An Introduction to Current Research (Wiley, New York, 1962)

[26] A. Ashtekar and R. O. Hansen, J. Math. Phys. 19 (1978) 1542

[27] R. Shoen and S. T. Yau, Commun. Math. Phys. 65 (1979) 45

[28] R. K. Sachs, Phys. Rev. 128 (1962) 2851

[29] R. Penrose, in: Group Theory in Non-Linear Problems (Ed. A. O. Barut, Reidel, Dordrecht 1974)

[30] E. T. Newman, Gen. Rel. and Grav. 7 (1976) 107

[31] M. Ko, E. T. Newman and K. P. Tod, in: Asymptotic Structure of Space-time (Eds. F.P. Esposito and L. Witten, Plenum, New York 1976)

[32] R. O. Hanson, E. T. Newman, R. Penrose and K. P. Tod, Proc. Roy. Soc. A363 (1978) 445

[33] R. Penrose, Gen. Rel. and Grav. 7 (1976) 31

[34] R. S. Ward, Phys. Lett 61A (1977) 81

[35] W. Rindler, Mon. Not. Roy. Astr. Soc. 116 (1956) 6

[36] R. Penrose, in: Battelle Recontres, 1967 Lectures in Mathematics and Physics (Eds. C. M. DeWitt and J. A. Wheeler, Benjamin New York 1968)

[37] R. Geroch, E. H. Kronheimer and R. Penrose, Proc. Roy. Soc. A327 (1972) 545

[38] B. Schmidt, Gen. Rel. and Grav. 1 (1971) 269

[39] B. Schmidt, Commun. Math. Phys. 36 (1974) 73

[40] B. Bosshardt, Commun. Math. Phys. 46 (1976) 263

[41] S. W. Hawking, Commun. Math. Phys. 25 (1972) 152

[42] R. Penrose in: Theoretical Principles in Astrophysics and Relativity (Eds. N. R. Lebovitz, W. H. Reid and P. O. Vandervoort, Chicago Univ. Press, Chicago 1978)

[43] R. Penrose and R. M. Floyd, Nature Physical Science 229 (1971) 117

[44] C. W. Misner, Phys. Rev. Lett. 28 (1972) 994

[45] B. Ya. Zel'dovich, Zh. Eksp. Teor. Fiz. 62 (1972) 2076; Engl. Transl.: Soviet Phys.-JETP 35 (1972) 1085

[46] J. D. Bekenstein, Phys. Rev. D7 (1973) 2333, D9 (1974) 3292

[47] S. W. Hawking, Commun. Math. Phys. 43 (1975) 199.

[48] M. J. Rees, Nature 266 (1977) 333

[49] R. Penrose, in: General Relativity, An Einstein Centenary Survey (Eds. S. W. Hawking and W. Israel, Cambridge Univ. Press, Cambridge 1979)

[50] B. J. Carr, Astrophys. J. 206 (1976) 8

[51] R. Geroch, J. Math. Phys. 11 (1970) 437

[52] C. J. Isham, R. Penrose and D. W. Sciama, Eds., Quantum Gravity, and Oxford Symposium (Clarendon Press, Oxford 1975)

113
TWISTORS FOR FLAT AND CURVED SPACE-TIME

General relativity and gravitation, Vol. 2, pp. 283–328, Plenum, New York and London, 1980. With R. S. Ward

A convenient twistor survey.

528 Twistors for flat and curved space-time. *General relativity and gravitation, Vol. 2,* pp. 283–328, Plenum, New York and London, 1980

Twistors for Flat and Curved Space-Time

R. PENROSE AND R. S. WARD

1. Introduction

Einstein's two great theories of relativity have each led to revolutionary (and presumably irreversible) changes in our conceptions of space and time. Indeed, the special theory now occupies so basic a role in physics that no attempt at progress towards a reformulation of theoretical physics can be taken very seriously which does not come profoundly to terms with it. And now also the general theory is just beginning to achieve an impressively secure experimental status. This is in addition to its already compelling stature as a logically satisfactory and esthetically pleasing theoretical edifice. So the much more exacting requirements of compatibility with that theory, in spirit as well as in detail, have also to be faced by any aspiring new formalism.

Twistor theory provides an attempt at such a formalism, whereby an alternative geometric framework for the description of basic physics is put foward. It is based, at the outset, on a reformulation of the geometry of special relativity. One of the prime motivations behind this reformulation is the thought that there should be a more intimate union between space–time structure and quantum mechanics than exists in conventional theory. For quantum theory is no less important than is relativity in governing the nature of physical processes. Now, a basic feature of quantum theory is that the complex field replaces the reals as the basic number system for describing the world. Quantum linear superposition is always carried out with complex coefficients, but this is not normally thought to have any intimate relation with the structure of space and time. In twistor theory, however, complex numbers also play a key role in determining the very structure of space–time.

R. PENROSE • Mathematical Institute, Oxford, England. R. S. WARD • Merton College, Oxford, England; present address: Trinity College, Dublin, Ireland.

In fact, the particular dimension and signature that our space–time possesses play special roles in the theory. The formalism is, in a clear sense, naturally tied to these special values of the space and time dimensions. And in a slightly indirect way, the same complex numbers may be said to govern both the space–time structure and the quantum-mechanical linear superposition.

As twistor theory stands, it is only Einstein's special relativity whose geometric structure is completely described, and Section 2 of this chapter will provide a short review of flat-space twistor theory. The various different roles that twistors can assume in special relativity will be given. We shall not, however, enter into an account of twistor particle theory since this will be covered in full detail in review articles by Hughston[1] and by Penrose and Sparling.[2]

The twistor description of general relativity is, as yet, only partially complete. The later sections of this chapter will give some indication of what is possible to achieve in this area, at present. We shall show (in Section 3) how the effect of gravitational impulsive waves can be described twistorially. Then we shall give an account of a method whereby all the self-dual (necessarily complex) solutions of Einstein's equations may be described in terms of a curved twistor space (Section 4.1). A large class of examples for which this procedure can be carried out explicitly will also be given (Section 4.2). When the space–time (Weyl) curvature is not self-dual, one method of proceeding is to refer the geometry to some specific hypersurface in the space–time. This gives rise to the concept of hypersurface twistor (Section 5.1). For an asymptotically flat space–time this hypersurface can be taken to be at (null) infinity and the concept of asymptotic twistor arises (Section 5.2). Asymptotic twistor space has precisely the structure of the twistor space for a self-dual vacuum complex space–time, which, in this case, turns out to be Newman's $\mathscr{H}$ space.[3]‡

2. Twistor Theory in Flat Space–Time

2.1. Twistor Algebra and Geometry

There are many essentially equivalent ways of setting up the correspondence between twistors and physical or geometrical structures in Minkowski space–time. Perhaps the most directly physical approach is to regard a twistor† (up to an undetermined phase multiplier) as representing the momentum and angular momentum structure of a classical zero-rest-mass free particle. Alternatively, a description of a twistor (up to a nonzero

† We refer here to the most basic type of twistor, namely, a flat-space twistor of valence $\left[{}^{1}_{0}\right]$. Those of general valence $\left[{}^{p}_{q}\right]$ are then obtained from these by the standard operations of taking duals and tensor products.

‡ See Chapter 6 by Boyer, Finley, and Plebański in this volume.

complex multiplier) can be given in terms of certain twisting congruences of null straight lines known as Robinson congruences. Yet again, a twistor may be viewed as a solution of the conformally invariant "twistor equation" or as such a solution paired with its derivative (which is more satisfactory for certain purposes). Or again, it is an element of the basic representation space of $SU(2,2)$, regarded as the fourfold cover of the identity-connected component of the conformal group of space–time. Or it may be viewed as a helicity-raising operator, taking massless free fields of helicity s into such fields of helicity $s + \frac{1}{2}\hbar$.

All these descriptions are concerned, essentially, with structures in *real* space–time. But it is also illuminating, and for our present purposes most useful, to give a basic description of a twistor in terms of *complexified* space–time geometry. This is the approach that we shall mainly follow here, but we shall also show how this viewpoint gives rise to the various others mentioned above.

Choose an origin O in (real) Minkowski space–time† $\mathbb{M}^I$ and let the point x be represented by the position vector x^a, whose components x^0, x^1, x^2, x^3 then give a standard Minkowski coordinate system for $\mathbb{M}^I$. These components can be reexpressed in 2-spinor terms as

$$\begin{bmatrix} x^{00'} & x^{01'} \\ x^{10'} & x^{11'} \end{bmatrix} = \frac{1}{2^{1/2}} \begin{bmatrix} x^0 + x^1 & x^2 + ix^3 \\ x^2 - ix^3 & x^0 - x^1 \end{bmatrix} \tag{2.1}$$

We shall adopt exactly the same correspondence when x^0, x^1, x^2, x^3 are allowed to be complex. Then x^a gives the position vector of a point x in *complexified* Minkowski space $\mathbb{CM}^I$. Thus x is a *real* point ($x \in \mathbb{M}^I$) if and only if $x^{AB'}$ is *Hermitian*.‡

The basic correspondence between $\mathbb{CM}^I$ and twistor space $\mathbb{T}$ is given by the fundamental relation

$$\omega^A = ix^{AA'}\pi_{A'} \tag{2.2}$$

which expresses the condition for the twistor $Z^\alpha \in \mathbb{T}$, given by $Z^\alpha = (\omega^A, \pi_{A'})$, to be *incident* with the point $x \in \mathbb{CM}^I$. The geometry of this correspondence shows up most clearly in terms of the *projective* twistor space $\mathbb{PT}$. The points of $\mathbb{PT}$ are the equivalence classes of (nonzero) proportional twistors, so each nonzero twistor $Z^\alpha \in \mathbb{T}$, corresponds to a *projective twistor*

$$\mathbf{Z} = \{\rho Z^\alpha | 0 \neq \rho \in \mathbb{C}\} \in \mathbb{PT}$$

† See page 328 for a list of the spaces and manifolds used in this chapter.

‡ Though it matters little at this stage, we adopt the "abstract index" conventions, whereby $x^a, x^{AB'}, g_{ab}, \xi_C$, etc. denote frame-independent (though sometimes origin-dependent) tensor or spinor (fields), while $x^a = (x^0, x^1, x^2, x^3)$, $x^{AB'}$, g_{ab}, ξ_C, etc. denote the corresponding sets of components. This enables us to write $x^a = x^{AA'}$, $g_{ab} = \varepsilon_{AB}\varepsilon_{A'B'}$, etc. To assert that $x^{AB'}$ is Hermitian is to assert that the matrix of standard components $x^{AB'}$ is Hermitian. The use of abstract indices does not play any important role here, however, until curved space–times are considered.

where $\rho Z^a = (\rho\omega^A, \rho\pi_{A'})$. The space **PT** is complex projective 3-space ($\mathbb{CP}^3$), and each of its points (i.e., each projective twistor) may be interpreted in $\mathbb{CM}^I$ as the set of complex space–time points incident with it, i.e., with the set of complex solutions of (2.2) in x, given when ω^A and $\pi_{A'}$ are held fixed. Inversely, each point of $\mathbb{CM}^I$ may be interpreted in **PT** as the set of projective twistors incident with it, i.e., with the set of solutions of (2.2) in Z^a, and, therefore, in **Z** [(2.2) being homogeneous] given, now, when x^a is held fixed.

It is clear from the form of (2.2) (since ω^A is just an arbitrary linear function of $\pi_{A'}$) that these latter sets are simply *the complex linear subspaces of* $\mathbb{T}$ *of dimension 2*, apart from those subspaces having nontrivial intersection with the 2-space I with equation $\pi_{A'} = 0$. In terms of **PT**, these are the *complex projective lines* in **PT**, apart from those meeting the line **I** (the projective version of I) (see Figure 1). Thus the points of complexified Minkowski space can be realized twistorially in this way. For the inverse construction we need to examine the loci in $\mathbb{CM}$ which are given by solving (2.2) for fixed $\omega^A, \pi_{A'}$. Now assume, first, that $\pi_{A'} \neq 0$ and let x_0^a be a particular solution, and x^a a general solution. Setting

$$n^a = x^a - x_0^a \tag{2.3}$$

we see that (2.2) gives

$$n^{AA'}\pi_{A'} = 0 \tag{2.4}$$

so n^a has the form

$$n^{AA'} = \lambda^A \pi^{A'} \tag{2.5}$$

(Raising and lowering of spinor indices is achieved by use of the skew-symmetric spinors $\varepsilon_{AB}, \varepsilon_{A'B'}, \varepsilon^{AB}, \varepsilon^{A'B'}$ whose standard components have the values given by $\mathbf{B} - \mathbf{A} = 0,\ 1,\ -1,\ 0 = \mathbf{B}' - \mathbf{A}'$. Thus $\xi^0 = \xi_1,\ \xi^1 =$

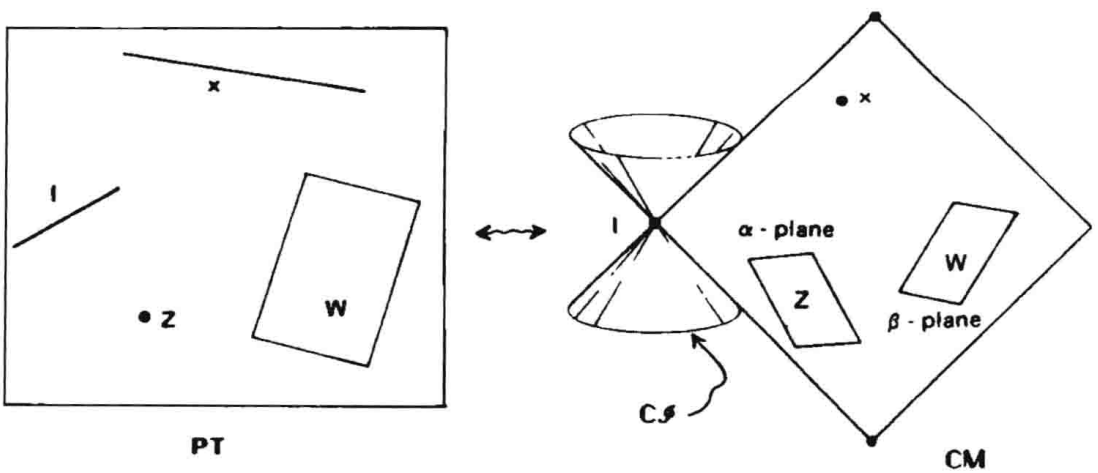


Fig. 1

$-\xi_0$, $\eta^{0'} = \eta_{1'}$, $\eta^{1'} = -\eta_{0'}$. By elementary lemmas, if ξ^A and $\eta^{A'}$ are non-zero, then $\psi_{A\cdots PQ}^{R'\cdots U'}\xi^Q = 0$ implies that $\psi_{A\cdots Q}^{R'\cdots U'}$ has the form $\phi_{A\cdots P}^{R'\cdots U'}\xi_Q$, whereas $\psi_{A\cdots Q}^{R'\cdots T'U'}\eta_{U'} = 0$ implies that it has the form $\chi_{A\cdots Q}^{R'\cdots T'}\eta^{U'}$, etc.) Allowing λ^A to vary in (2.5) we obtain a two-complex-parameter family of mutually orthogonal complex null vectors n^a [since $(\lambda^A\pi^{A'})(\tilde{\lambda}_A\pi_{A'}) = 0$]. It follows that the set $\mathcal{Z}$ of solutions of (2.2) is a *totally null complex 2-plane*, i.e., a 2-plane all of whose tangent vectors are null and mutually orthogonal, i.e., whose metric induced from the complex metric g_{ab} vanishes.

Totally null 2-planes in $\mathbb{CM}^I$ are of two distinct kinds, those that we have just been discussing, called *α-planes*, with tangent vectors of the form (2.5) where $\pi^{A'}$ is held fixed and λ^A varies, and those, called *β-planes* having tangent vectors of the form (2.5) with λ^A fixed and $\pi^{A'}$ varying. Thus points of $\mathbb{PT}$ (at least, those not on $\mathbf{I}$) are represented in $\mathbb{CM}^I$ by α-planes. (The interpretation of the β-planes will be given in a moment.)

To interpret the points of $\mathbf{I}$ ($\pi_{A'} = 0$) we need to pass to the *compactified* complex Minkowski space $\mathbb{CM}$ whose points may be conveniently represented as the lines of $\mathbb{PT}$ *including* those which meet $\mathbf{I}$. Thus, $\mathbb{CM}$ is composed not only of the points of $\mathbb{CM}^I$ (lines of $\mathbb{PT}$ not meeting $\mathbf{I}$) but, in addition, the points of a complex 3-surface $\mathbb{CS}$ representing the lines of $\mathbb{PT}$ which do meet $\mathbf{I}$. The 3-surface $\mathbb{CS}$ is null (defining the "points at infinity" for $\mathbb{CM}^I$) and is a *cone*, whose vertex I is represented, in $\mathbb{PT}$, by the line $\mathbf{I}$ itself. The entire complex 4-manifold $\mathbb{CM}$ is *compact* and can, in fact, be realized[4] as a nonsingular quadric hypersurface in complex projective 5-space ($\mathbb{CP}^5$). This is known as the *Klein* (or Grassmann) representation of lines in $\mathbb{CP}^3$. The space $\mathbb{CM}$ has a well-defined complex conformal metric, defining a complex light cone at each of its points. This light cone can be pictured, in $\mathbb{CP}^5$, as the intersection of the quadric with the hyperplane touching it at the point in question. The null lines of $\mathbb{CM}$ are actually the *straight lines* on the quadric, the α-planes and β-planes constituting the two families of complex 2-planes lying on the quadric.

Though the point I and the hypersurface $\mathbb{CS}$ define "infinity" for the metric geometry of $\mathbb{CM}$, they have no special relation to its conformal geometry, I being on an equal conformal footing with all other points of $\mathbb{CM}$. The linear transformations of $\mathbb{T}$ form a 16-complex-parameter group [$GL(4, \mathbb{C})$], which induces the 15-complex-parameter projective linear group on $\mathbb{PT}$. In terms of $\mathbb{CM}$, this projective group becomes the 15-complex-parameter (proper) conformal group $C(4, \mathbb{C})$, expressing the (orientation-preserving) symmetries of complexified compactified Minkowski space (the complex light cones being preserved by the transformations). The (proper) complexified Poincaré group, together with dilations, arises as the subgroup of this group for which the line $\mathbf{I}$ of $\mathbb{PT}$ is left invariant. (To eliminate the dilations a further restriction is needed. We shall discuss this later.)

Next, in order to build up twistor algebra, we need the concept of a *dual twistor* W_α. We say that $W_\alpha = (\lambda_A, \mu^{A'})$, is *incident with the point* $x \in \mathbb{CM}^I$ whenever the fundamental relation

$$\mu^{A'} = -i\lambda_A x^{AA'} \tag{2.6}$$

holds, and it is *incident with the twistor* $Z^\alpha = (\omega^A, \pi_{A'})$ whenever the scalar product

$$Z^\alpha W_\alpha \equiv \omega^A \lambda_A + \pi_{A'} \mu^{A'} \tag{2.7}$$

vanishes. Note that the scalar product (2.7) indeed vanishes whenever the relations (2.6) and (2.2) both hold together, i.e., whenever there is a point of $\mathbb{CM}^I$ incident with both. Provided that $W_\alpha \neq 0$, the *projective dual twistor* $\mathbf{W} = \{\rho W_\alpha | 0 \neq \rho \in \mathbb{C}\}$ is defined, and is represented in $\mathbb{PT}$ by the system of projective twistors incident with it, i.e., by the projective plane in $\mathbb{PT}$ whose equation, in Z^α, is $Z^\alpha W_\alpha = 0$. In terms of $\mathbb{T}$, this is a complex linear subspace of dimension 3. In terms of $\mathbb{CM}^I$, $\mathbf{W}$ is represented as the set of complex solutions of (2.6) given when $\mu^{A'}$ and λ_A are held fixed, i.e., according to our previous discussion, by a β-plane. It follows that β-planes in $\mathbb{CM}$ are represented by complex projective planes in $\mathbb{PT}$. An α-plane is incident with a β-plane if the two intersect in $\mathbb{CM}$ (possibly on $\mathbb{C}\mathscr{I}$); this occurs whenever the point of $\mathbb{PT}$, representing the α-plane, lies on the plane in $\mathbb{PT}$, representing the β-plane. A point of $\mathbb{CM}$ is incident with an α-plane or, alternatively, with a β-plane if it lies on it; equivalently, this is the respective condition that the corresponding line in $\mathbb{PT}$ contain the point representing the α-plane or lie on the plane representing the β-plane. We note that if a point in $\mathbb{PT}$ lies on a plane in $\mathbb{PT}$ then there is a one-parameter family of lines lying in the plane and containing the point. Otherwise there are no such lines at all. Such a one-parameter family of lines is called a *plane pencil*. Thus, if an α-plane and a β-plane do intersect in $\mathbb{CM}$ (i.e., are incident), they have a (complex) curve in common. This curve is easily seen, from (2.5), to be a complex null geodesic (i.e., null straight line) in $\mathbb{CM}$, so null geodesics in $\mathbb{CM}$ are represented in $\mathbb{PT}$ by plane pencils.

We note that the *dual* twistor space $\mathbb{T}^*$, and its projective version $\mathbb{PT}^*$, can also be used to describe the geometry of $\mathbb{CM}$. The only difference is that the roles of points and planes in $\mathbb{PT}$ are interchanged in $\mathbb{PT}^*$. Thus points in $\mathbb{PT}^*$ now represent β-planes in $\mathbb{CM}$ and planes in $\mathbb{PT}^*$ represent α-planes in $\mathbb{CM}$. As before, projective lines and plane pencils in $\mathbb{PT}^*$, respectively, represent points and null geodesics in $\mathbb{CM}$.

So far, we have paid no attention to the *real* structure of $\mathbb{CM}$, i.e., to the particular subset $\mathbb{M}$ of $\mathbb{CM}$. Another way of specifying this real structure is in terms of the operation $\mathscr{C}$, which sends points of $\mathbb{CM}$ into their complex conjugates. Comparing (2.2) with (2.6) we observe that complex conjugation interchanges twistors with dual twistors, and α-planes with β-

planes. The complex conjugate of the twistor $Z^\alpha = (\omega^A, \pi_{A'})$ is the dual twistor $\bar{Z}_\alpha = (\bar{\pi}_A, \bar{\omega}^{A'})$ and, correspondingly, the complex conjugate of the dual twistor $W_\alpha = (\lambda_A, \mu^{A'})$ is the twistor $\bar{W}^\alpha = (\bar{\mu}^A, \bar{\lambda}_{A'})$. This enables the twistor *norm* of Z^α and of W_α to be defined as, respectively,

$$Z^\alpha \bar{Z}_\alpha = \omega^A \bar{\pi}_A + \pi_{A'} \bar{\omega}^A \tag{2.8}$$

and

$$W_\alpha \bar{W}^\alpha = \lambda_A \bar{\mu}^A + \mu^{A'} \bar{\lambda}_A \tag{2.9}$$

This norm is a Hermitian form of signature $++--$, so the linear transformations of $\mathbb{T}$ which preserve it (or, equivalently, which preserve the relation between a twistor and its complex conjugate) constitute a realization of the 16-real-parameter group $U(2, 2)$. These transformations induce maps of the *real* compactified Minkowski space[5] $\mathbb{M}$ to itself, preserving its conformal structure. But among the elements of $U(2, 2)$ are the phase transformations

$$Z^\alpha \mapsto e^{i\theta} Z^\alpha \qquad (\theta \text{ real}) \tag{2.10}$$

which yield the identity on $\mathbb{PT}$ and, consequently, the identity on $\mathbb{M}$ also. The factor group of $U(2, 2)$ by the transformations (2.10) is a 15-real-parameter group isomorphic with the connected component $C^\uparrow_+(1, 3)$ of the *conformal group* $C(1, 3)$ on $\mathbb{M}$. Instead of considering this factor group, however, we may prefer to normalize the elements of $U(2, 2)$ by requiring them to have determinant unity, i.e., to belong to the subgroup $SU(2, 2)$. This cuts down the freedom in (2.10) to the discrete ($\mathbb{Z}_4$) subgroup whose elements are the four transformations

$$Z^\alpha \mapsto Z^\alpha, \qquad Z^\alpha \mapsto iZ^\alpha, \qquad Z^\alpha \mapsto -Z^\alpha, \qquad Z^\alpha \mapsto -iZ^\alpha \tag{2.11}$$

these being precisely the elements of $SU(2, 2)$ that induce the identity $C(1, 3)$ transformation on $\mathbb{M}$. The group $SU(2, 2)$ is thus seen to be 4 : 1 isomorphic with $C^\uparrow_+(1, 3)$, this fourfold ambiguity in the passage from $C^\uparrow_+(1, 3)$ to $SU(2, 2)$ being *essential*, owing to the fact that $SU(2, 2)$ is a connected group. (This has the implication that any continuous, local, conformally invariant geometrical realization of a twistor Z^α in space–time terms, in $\mathbb{M}$, must be ambiguous in that iZ^α, $-Z^\alpha$, and $-iZ^\alpha$ have the same realization as Z^α.)

A basic twistor Z^α is now seen to be interpretable as an element of the representation space $\mathbb{T}$, on which $SU(2, 2)$ acts linearly, preserving the norm (2.8), where $SU(2, 2)$ is regarded as the fourfold cover of the group $C^\uparrow_+(1, 3)$ acting on $\mathbb{M}$. Other finite-dimensional representations are obtained by taking the corresponding action on the dual space $\mathbb{T}^*$ or on the tensor

product of p copies of $\mathbb{T}$ with q copies of $\mathbb{T}^*$:

$$\underbrace{\mathbb{T}\otimes\cdots\otimes\mathbb{T}}_{p}\otimes\underbrace{\mathbb{T}^*\otimes\cdots\otimes\mathbb{T}^*}_{q}$$

An element of this space—or the abstract-index form of such an element—is called a *twistor of valence* $\left[\begin{smallmatrix}p\\q\end{smallmatrix}\right]$, so the original basic twistors have valence $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ and their duals have valence $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$. The complex conjugate of a twistor of valence $\left[\begin{smallmatrix}p\\q\end{smallmatrix}\right]$ is a twistor of valence $\left[\begin{smallmatrix}q\\p\end{smallmatrix}\right]$. The standard notation $T^{\alpha_1\cdots\alpha_p}_{\beta_1\cdots\beta_q}$ is used for a twistor of valence $\left[\begin{smallmatrix}p\\q\end{smallmatrix}\right]$ and $\bar{T}^{\beta_1\cdots\beta_q}_{\alpha_1\cdots\alpha_p}$ for its complex conjugate, of valence $\left[\begin{smallmatrix}q\\p\end{smallmatrix}\right]$. The standard tensor notation for sums, outer products, contractions, index substitutions, symmetries (with round and square brackets denoting symmetric and antisymmetric parts, respectively) then applies generally to twistors.

A twistor Z^α (or, equivalently, $\bar{Z}_\alpha$) is said to be *positive* (right handed), *negative* (left handed), or *null* according as $Z^\alpha\bar{Z}_\alpha$ is positive, negative, or zero. This characterization clearly also applies to the projective twistor $\mathbf{Z}$ (or to $\bar{\mathbf{Z}}$). Thus $\mathbb{T}$ and $\mathbb{PT}$ are each divided into three portions, labeled $\mathbb{T}^+$, $\mathbb{T}^-$, and $\mathbb{N}$, and correspondingly, $\mathbb{PT}^+$, $\mathbb{PT}^-$, and $\mathbb{PN}$, where the elements of $\mathbb{T}^+$ are the positive twistors, $\mathbb{T}^-$ the negative twistors, and $\mathbb{N}$ the null twistors. The topology of each of the complex 3-manifolds $\mathbb{PT}^+$ and $\mathbb{PT}^-$ is $S^2\times\mathbb{R}^4$ and of each of the complex 4-manifolds $\mathbb{T}^+$ and $\mathbb{T}^-$ is $S^3\times\mathbb{R}^5$; the topology of the real 5-manifold $\mathbb{PN}$ is $S^2\times S^3$ and of the real 7-manifold $\mathbb{N}$ (where the origin $Z^\alpha=0$ is here excluded) is $S^3\times S^3\times\mathbb{R}$.

The elements of $\mathbb{PN}$ are very simply interpreted in Minkowski space terms as *light rays* (i.e., as null straight lines), where we include under this heading the generators of $\mathscr{I}$. (The light cone at infinity $\mathscr{I}$, for $\mathbf{M}$, is the set of real points of $\mathbb{C}\mathscr{I}$ and it is generated by an S^2 family of such light rays.[6]) For $Z^\alpha\bar{Z}_\alpha=0$ is the condition that the α-plane defined by $\mathbf{Z}$ and the β-plane defined by $\bar{\mathbf{Z}}$ be incident, and we have seen that when an α-plane meets a β-plane they have a complex null geodesic in common. This complex null geodesic must here contain a *real* null geodesic, since complex conjugation interchanges the α-plane with the β-plane and so leaves their intersection invariant. This real null geodesic (or light ray) is the intersection of the α-plane with $\mathbb{M}$ and also of the β-plane with $\mathbf{M}$ (i.e., it is precisely the set of real points of each of these complex null planes) and it determines both planes uniquely. These properties can also be directly seen by referring back to (2.2), in the case when $\pi_{A'}\neq 0$. Thus the twistor $Z^\alpha=(ix^{AA'}\pi_{A'},\pi_{A'})$ (with $\bar{x}^a=x^a$) describes the light ray through the real point with position vector x^a, in the direction of $\bar{\pi}^A\pi^{A'}$.

When $Z^\alpha\bar{Z}_\alpha\neq 0$, an interpretation of $\mathbf{Z}$ in real space–time terms can still be given, but is more complicated. We can interpret $\mathbf{Z}$ in terms of the family of null dual projective twistors $\mathbf{W}$ that are incident with it. Let x^a be

real, so that $W_\alpha = (\lambda_A, -i\lambda_A x^{AA'})$ represents the light ray through x^a in the direction of $\lambda^A \bar{\lambda}^{A'}$. For W_α to be incident with $Z^\alpha = (\omega^A, \pi_{A'})$, we have

$$\begin{aligned} 0 &= W_\alpha Z^\alpha \\ &= \lambda_A \omega^A - i\lambda_A x^{AA'} \pi_{A'} \\ &= \lambda_A (\omega^A - i x^{AA'} \pi_{A'}) \end{aligned} \tag{2.12}$$

that is, λ^A is proportional to $\omega^A - ix^{AA'}\pi_{A'}$. (Note that $\omega^A \neq ix^{AA'}\pi_{A'}$ for real x^a if $Z^\alpha \bar{Z}_\alpha \neq 0$.) Thus, the direction of the W light ray through the point x of $\mathbb{M}^I$ is the null direction defined by the spinor

$$\omega^A(x) = \omega^A - ix^{AA'}\pi_{A'} \tag{2.13}$$

As x^a varies, (2.13) provides us with a spinor field, the null directions defined by $\omega^A(x)$ being tangential to the required family of W light rays. Such a family, defined by (2.13), is called a *Robinson congruence* (where "congruence" means, in this context, a three-parameter family of lines). A Robinson congruence is necessarily shear free.[5] It twists consistently in a right-handed, or left-handed, sense at every point according as $Z^\alpha \bar{Z}_\alpha$ is positive or negative.

If we specialize to the case $Z^\alpha \bar{Z}_\alpha = 0$ (with $\pi_{A'} \neq 0$) we see that for those points x on the light ray Z defined by $\mathbf{Z}$ (i.e., for real x, incident with $\mathbf{Z}$), we have $\omega^A(x) = 0$, so *every* $\mathbf{W}$ incident with such an x satisfies (2.12). That is to say, the congruence of W rays has become a *special* Robinson congruence, i.e., simply the family of light rays meeting the particular light ray defined by $\mathbf{Z}$. (Recall, also, that $Z^\alpha W_\alpha = 0$ is the condition for the α-plane Z to meet the β-plane W; for *null* Z^α, W_α this is the condition for the light ray Z to meet the light ray W, possibly at $\mathscr{I}$.) If $\pi_{A'} = 0$, (2.13) defines a congruence of *parallel* light rays, these being the rays meeting the particular generator Z of $\mathscr{I}$.

In the above description we have been thinking of a Robinson congruence in terms of a family of projective dual twistors, and it is convenient to picture this in the space $\mathbb{PT}^*$. (A description in terms of $\mathbb{PT}$ can, of course, equally well be given, where we simply take complex conjugates throughout.) Thus $\mathbf{Z}$ is now represented as a plane in $\mathbb{PT}^*$, and the family of $\mathbf{W}$'s incident with $\mathbf{Z}$ by the system of points on this plane. The *null* dual projective twistors are the points of $\mathbb{PN}^*$, so the Robinson congruence defined by $\mathbf{Z}$ is described, in $\mathbb{PT}^*$, by those $\mathbf{W}$'s lying on the intersection of $\mathbb{PN}^*$ with the plane $\mathbf{Z}$ (Figure 2). The plane $\mathbf{Z}$ is four-real-dimensional, while $\mathbb{PN}^*$ is a five-real-dimensional hypersurface in $\mathbb{PT}^*$. Thus the intersection of $\mathbb{PN}^*$ with $\mathbf{Z}$ is three-real-dimensional and defines a congruence in $\mathbb{M}$, as required. The whole of $\mathbb{PN}^*$ may be thought of as representing the totality of *all* light rays in $\mathbb{M}$ (as, equally well, $\mathbb{PN}$ may be so

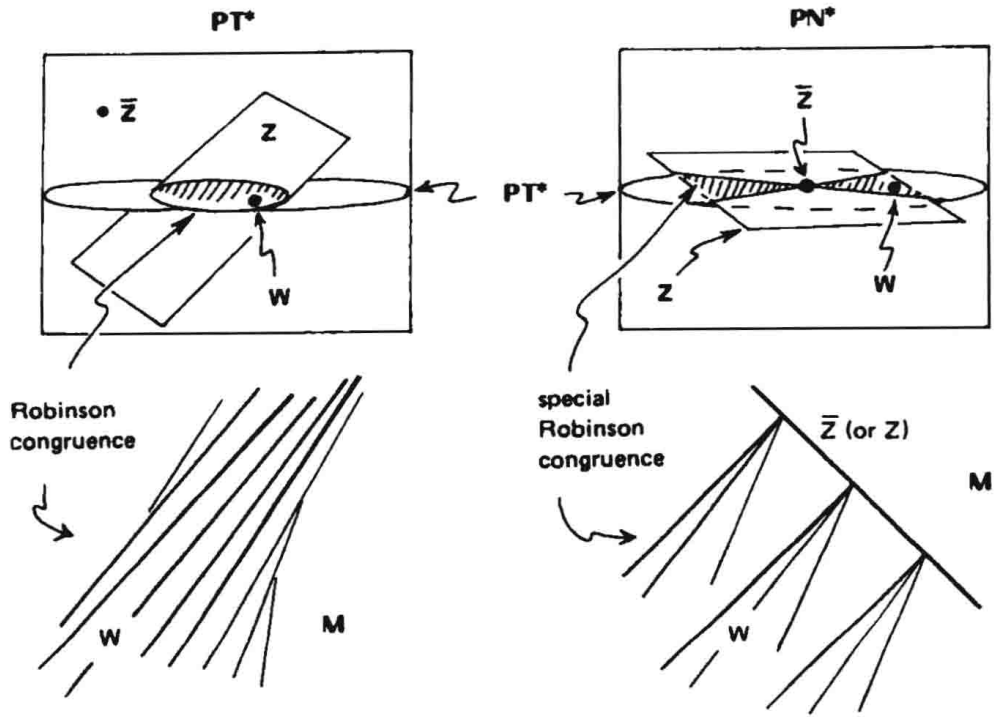


Fig. 2

viewed); this three-dimensional subset picks out the required Robinson congruence.

The projective dual twistor $\bar{\mathbf{Z}}$ is a point in $\mathbb{PT}^*$. It lies on $\mathbb{PN}^*$ if and only if it lies on the plane $\mathbf{Z}$, which is the condition for the Robinson congruence to become special. In this case, $\bar{\mathbf{Z}}$ describes a particular light ray in $\mathbb{M}$ (i.e., the same light ray as is defined by $\mathbf{Z}$) and the congruence becomes the family of light rays meeting it. Another way of describing such a congruence is as the system of generators of the various light cones in $\mathbb{M}$ whose vertices are the points of the particular light ray Z. Each such light cone is described, in $\mathbb{PT}^*$, by the points of a projective line in $\mathbb{PN}^*$ through the point $\bar{\mathbf{Z}}$. Thus, in this case, the intersection of $\mathbb{PN}^*$ with the plane $\mathbf{Z}$ is a *cone*, with vertex $\bar{\mathbf{Z}}$, and whose generators are projective lines (topology S^2) belonging to a one-real-parameter (S^1) family.

Let us examine a little more closely the twistor description of light cones in $\mathbb{M}$, where, for convenience, we now revert to $\mathbb{PT}$, rather than the equivalent $\mathbb{PT}^*$. If X^α and Y^α are (nonzero) null twistors, representing distinct light rays in $\mathbb{M}$ then, as we have seen, the twistor orthogonality condition $X^\alpha \bar{Y}_\alpha = 0$ is necessary and sufficient for the light rays X and Y to meet (possibly at $\mathscr{I}$). When this condition holds, every twistor Z^α, defined by

$$Z^\alpha = \rho X^\alpha + \sigma Y^\alpha \tag{2.14}$$

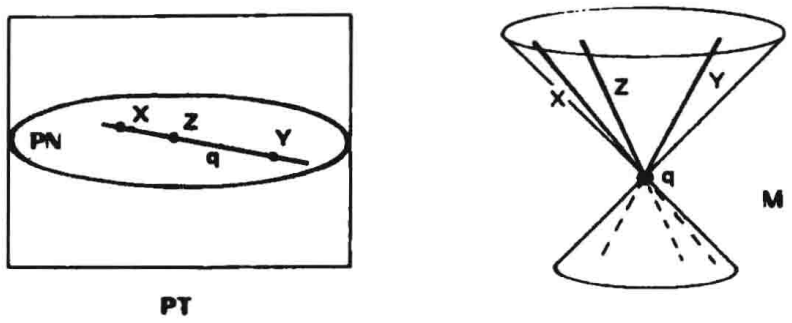


Fig. 3

(with ρ, σ not both zero), is also null (and nonzero) and is orthogonal to each of X^a and Y^a. Thus it represents a light ray meeting each of the null lines X, Y, i.e., containing their intersection point q. Thus, as the ratio $\rho:\sigma$ varies, the projective twistors $\mathbf{Z}$, defined by (2.14), give the various different light rays generating the light cone of q (Figure 3). In terms of $\mathbb{PT}$, $\mathbf{X}$ and $\mathbf{Y}$ are two points of $\mathbb{PN}$, the orthogonality condition $X^a\bar{Y}_a = 0$ having the interpretation that the projective line joining $\mathbf{X}$ and $\mathbf{Y}$ lies entirely in $\mathbb{PN}$. The various points $\mathbf{Z}$ of this line represent the different generators of the light cone of q; indeed, the line itself represents the point q in accordance with the correspondence described earlier, whereby points of $\mathbb{CM}$ are represented by projective lines in $\mathbb{PT}$. In this particular case, q is a *real* point, which is interpreted, in $\mathbb{PT}$, as the fact that the corresponding projective line lies entirely in $\mathbb{PN}$. We note that the topology of any projective line in $\mathbb{PT}$ is S^2, since the space of ratios $\rho:\sigma$ of a pair of complex numbers is the Riemann sphere (Argand plane together with a point at infinity). For a real point q, this sphere is realized as the space of light rays through q, i.e., as the "sky" or "celestial sphere" of an observer at q.

Let us now return to consideration of the spinor field $\omega^A(x)$ given by (2.13). We note that it satisfies the equation

$$\nabla^{(A}_{A'}\omega^{B)}(x) = 0 \tag{2.15}$$

referred to as the *twistor equation*, where ∇_a is the standard covariant derivative operator ($\nabla_a = \partial/\partial z^a$ in Minkowski coordinates) for the flat-space $\mathbb{M}^I$. Conversely *every* solution of (2.15) has the form (2.13). [To see this, use the commutativity of derivative operators and equation (2.15) to show that $\nabla^A_{A'}\nabla^B_{B'}\omega^C(x) = \nabla^{[A}_{A'}\nabla^B_{B'}\omega^{C]}(x) = 0$, so $\omega^A(x)$ is linear in x^a and, by (2.15) again, has the required form (2.13).] The (constant) spinors ω^A and $\pi_{A'}$ are determined, respectively, by the value of the field $\omega^A(x)$ and its first derivative, at the origin $x^a = 0$. Thus we have an alternative way of viewing a twistor, namely, as a solution of (2.15). These solutions form a four-complex-dimensional vector space which may be identified with $\mathbb{T}$. The

zero set of each solution, i.e., the set of x's, now allowed to be *complex*, at which $\omega^{A}(x) = 0$, defines the α-plane of the twistor. But now the scaling is fixed so that the actual twistor Z^{α}, and not just the projective twistor $\mathbf{Z}$ is represented.† This viewpoint has a certain advantage in that the transformation properties of twistors under conformal transformations become much more evident as, indeed, does the very fact that the twistor concept is a conformally invariant one. This is because (2.15) is an invariant equation under conformal rescalings of the metric.[7]

We shall not be particularly concerned here with questions of conformal invariance. However, the dependence on position provided by $\omega^{A}(x)$ is important also for Poincaré invariance. We may regard $\omega^{A}(x)$, in fact, as the "ω part" of the twistor Z^{α}, *when referred to the real point x as origin*, the "π part" of Z^{α} being taken to be origin independent, i.e.,

$$\pi_{A'}(x) = \pi_{A'} \tag{2.16}$$

The compatibility of this with the fundamental relation (2.2) follows directly. For let y^{a} be the position vector of $y \in \mathbb{CM}^{I}$ with respect to the standard origin, so that with respect to x as origin the position vector of y is

$$y^{a}(x) = y^{a} - x^{a} \tag{2.17}$$

Thus by (2.13), (2.16), and (2.17)

$$\omega^{A}(x) - iy^{AA'}(x)\pi_{A'}(x) = \omega^{A} - iy^{AA'}\pi_{A'} \tag{2.18}$$

so that the condition that Z^{α} be incident with y^{a} has the same expression whichever origin is used. We note, also, that the *twistor norm* (2.8) is invariant under (2.13), (2.16) (x^{a} real) and is therefore Poincaré invariant in this sense. The description of a twistor Z^{α} as a pair of spinors $(\omega^{A}, \pi_{A'})$ is more satisfactorily replaced, from the point of view of Poincaré invariance, by one in which the origin dependence is made explicit by writing, instead

$$\begin{aligned} Z^{\alpha} &= (\omega^{A}(x), \pi_{A'}(x)) \\ &= (\omega^{A} - ix^{AA'}\pi_{A'}, \pi_{A'}) \end{aligned} \tag{2.19}$$

Such a description of a twistor as a pair of spinor *fields* (the second of which is constant and representing, essentially, the derivative of the first) incorporates both the translational and Lorentz-rotational behavior of the twistor. For simplicity, however, the original description $Z^{\alpha} = (\omega^{A}, \pi_{A'})$ will usually be adopted here.

It may be remarked, at this juncture, that the constancy (2.16), of $\pi_{A'}(x)$ can be expressed in another way, namely, as the fact that there is a Poincaré

† The question of the role of the ambiguity (2.11) has not been entered into here. A complete discussion requires an examination of the behavior of the solutions of (2.15) across $\mathscr{I}$. Here we consider only $\mathbb{M}^{I}$ (or $\mathbb{CM}^{I}$).

invariant projection map from $\mathbb{T}$ to the space $\tilde{\mathbb{S}}^*$ of primed "covariant" (i.e., lower-index) spinors $\pi_{A'}$. Likewise, there is a Poincaré-invariant injection of the space $\mathbb{S}$ of unprimed "contravariant" (i.e., upper-index) spinors ω^A into $\mathbb{T}$, completing the *exact sequence*[11] of maps between complex vector spaces:

$$0 \to \mathbb{S} \to \mathbb{T} \to \tilde{\mathbb{S}}^* \to 0 \tag{2.20}$$

The second map (injection) simply takes ω^A to $(\omega^A, 0)$ [which is Poincaré invariant because, now, $\pi_{A'} = 0$ in (2.19)] and the third (projection) takes $(\omega^A, \pi_{A'})$ to $\pi_{A'}$, as remarked before. The exactness is obvious. The dual sequence to (2.20)

$$0 \leftarrow \mathbb{S}^* \leftarrow \mathbb{T}^* \leftarrow \tilde{\mathbb{S}} \leftarrow 0 \tag{2.21}$$

is also (reading it now in the reverse direction) the complex conjugate of (2.20).

The linear transformations of $\mathbb{T}$ that preserve the sequence (2.20) and its complex conjugate relation to (2.21) give the (restricted) Poincaré group action (in fact twofold covered) on $\mathbb{M}$. In this sense, the sequence (2.20) contains the essence of the Poincaré-invariant structure of $\mathbb{T}$. Another way of expressing this Poincaré invariant structure is in terms of the *infinity twistors*:

$$I^{\alpha\beta} = \begin{bmatrix} \varepsilon^{AB} & 0 \\ 0 & 0 \end{bmatrix}, \qquad I_{\alpha\beta} = \begin{bmatrix} 0 & 0 \\ 0 & \varepsilon^{A'B'} \end{bmatrix} \tag{2.22}$$

which are antisymmetric and satisfy

$$\begin{gathered} I^{\alpha\beta}I_{\beta\gamma} = 0, \qquad I^{[\alpha\beta}I^{\gamma]\delta} = 0, \qquad I_{[\alpha\beta}I_{\gamma]\delta} = 0 \\ I_{\alpha\beta} = \tfrac{1}{2}I^{\gamma\delta}\varepsilon_{\alpha\beta\gamma\delta}, \qquad I^{\alpha\beta} = \tfrac{1}{2}I_{\gamma\delta}\varepsilon^{\alpha\beta\gamma\delta} \end{gathered} \tag{2.23}$$

where $\varepsilon_{\alpha\beta\gamma\delta}$ and $\varepsilon^{\alpha\beta\gamma\delta}$ are the antisymmetric Levi-Cività twistors, with $\varepsilon_{0123} = \varepsilon^{0123} = 1$. The $SU(2,2)$ transformations leave $\varepsilon_{\alpha\beta\gamma\delta}$ and $\varepsilon^{\alpha\beta\gamma\delta}$ invariant; the complex Poincaré group arises when $I^{\alpha\beta}$ and $I_{\alpha\beta}$ are specified as invariant (whereby $\varepsilon_{\alpha\beta\gamma\delta}$ and $\varepsilon^{\alpha\beta\gamma\delta}$ are consequently also invariant), the real Poincaré group preserving also the operation of twistor complex conjugation. The projection $\mathbb{T} \to \tilde{\mathbb{S}}^*$ is achieved, in essence, by the map $Z^\alpha \mapsto I_{\alpha\beta}Z^\beta$, while the injection $\mathbb{S} \to \mathbb{T}$, in essence, interprets ω^A as a twistor of the form $Z^\alpha = I^{\alpha\beta}U_\beta$. Thus the second and third maps of the sequence (2.20) are closely associated with, respectively, $I^{\alpha\beta}$ and $I_{\alpha\beta}$.

The twistors $I^{\alpha\beta}$ and $I_{\alpha\beta}$ represent the vertex I, of $\mathscr{I}$, in $\mathbb{M}$ (or $\mathbb{CM}$) or, equivalently, the line $\mathbf{I}$ in $\mathbb{PT}$ (which in fact lies in $\mathbb{PN}$). This is a special case of a general way that a point R of $\mathbb{CM}$—i.e., a line $\mathbf{R}$ in $\mathbb{PT}$—may be represented by a simple antisymmetric twistor $R^{\alpha\beta}$ or by its dual $R_{\alpha\beta}$:

$$R_{\alpha\beta} = \tfrac{1}{2}R^{\gamma\delta}\varepsilon_{\alpha\beta\gamma\delta}, \qquad R^{\alpha\beta} = \tfrac{1}{2}R_{\gamma\delta}\varepsilon^{\alpha\beta\gamma\delta} \tag{2.24}$$

For we have

$$R^{\alpha\beta} = X^{[\alpha}Y^{\beta]} \qquad \text{and} \qquad R_{\alpha\beta} = U_{[\alpha}V_{\beta]}$$

X and **Y** being two distinct points on the line **R** in $\mathbb{PT}$, and **U** and **V** two distinct planes through it. For R not on $\mathscr{I}$ we can adopt the convenient normalization $R^{\alpha\beta}I_{\alpha\beta} = 2 = R_{\alpha\beta}I^{\alpha\beta}$ and obtain[7]

$$R^{\alpha\beta} = \begin{bmatrix} -\frac{1}{2}\varepsilon^{AB}r_{c'}r^{c'} & ir^{A}{}_{B'} \\ -ir_{A'}{}^{B} & \varepsilon_{A'B'} \end{bmatrix}, \qquad R_{\alpha\beta} = \begin{bmatrix} \varepsilon_{AB} & ir_{A}{}^{B'} \\ -ir^{A'}{}_{B} & -\frac{1}{2}\varepsilon^{A'B'}r_{c}r^{c} \end{bmatrix} \tag{2.25}$$

where r^a is the position vector of the point R in $\mathbb{CM}^I$.

We have seen that a null twistor Z^α defines a light ray in $\mathbb{M}$, but have not yet given a physical description of a *non*-null twistor. In fact non-null twistors turn out to be closely associated with free massless particles with *spin*. The dependence, (2.19), that a twistor has on position (i.e., on choice of "origin") is, indeed, intimately related to the way that angular momentum depends on the choice of origin. For if $M^{ab}(=-M^{ba})$ and p_a describe the angular momentum and momentum, respectively, of some finite special-relativistic system in $\mathbb{M}^I$, referred to some standard origin O, and if $M^{ab}(x)$ and $p_a(x)$ are the corresponding quantities referred to the point x as origin, with position vector x^a with respect to O, then we have

$$M^{ab}(x) = M^{ab} - x^a p^b + x^b p^a \tag{2.26}$$

and

$$p_a(x) = p_a$$

Now, observe that if we set

$$\begin{aligned} M^{AA'BB'}(x) &= i\omega^{(A}(x)\bar{\pi}^{B)}(x)\varepsilon^{A'B'} - i\varepsilon^{AB}\bar{\omega}^{(A'}(x)\pi^{B')}(x) \\ p_{AA'}(x) &= \bar{\pi}_A(x)\pi_{A'}(x) \end{aligned} \tag{2.27}$$

where $Z^\alpha = (\omega^A(x), \pi_{A'}(x))$ as in (2.19), then (2.26) follows automatically. Thus the twistor Z^α defines, in a Poincaré-invariant way, the angular momentum and momentum—or, in short, the *kinematics*—of some special-relativistic system. It is not the most general such system, however, because there are ten free real parameters required for the kinematics (six for M^{ab} and four for p_a), whereas Z^α contains only the information of eight real parameters. Furthermore, the phase transformations

$$Z^\alpha \mapsto e^{i\theta}Z^\alpha \qquad (\theta \text{ real}) \tag{2.28}$$

leave (2.27) invariant, so that (at most) a seven-real-parameter family of kinematic structures are obtainable by means of (2.27). But it turns out, somewhat remarkably, that those kinematic structures that *are* so obtainable (with $\pi_{A'} \neq 0$) are precisely those defined by a *spinning massless particle*.

The requirements for the kinematics of such a particle are, first, that the momentum p_a be null and future pointing,

$$p_a p^a = 0, \qquad p_0 > 0 \tag{2.29}$$

and, second, that the Pauli–Lubanski spin vector

$$S_a = \tfrac{1}{2} e_{abcd} p^b M^{cd} \tag{2.30}$$

(where e_{abcd} is the antisymmetric Levi-Cività tensor, with $e_{0123} = 1 = -e^{0123}$) be a scalar multiple of the momentum

$$S_a = s p_a \tag{2.31}$$

Defining M^{ab} and p_a by (2.27) (with $\pi_{A'} \neq 0$), where by the preceding discussion it is sufficient just to consider $x^a = 0$, we find (2.29) and (2.31) to be satisfied identically. The verification for (2.29) is trivial, whereas that for (2.31) follows directly from the spinor formula[7]

$$e_{AA'BB'CC'DD'} = i\varepsilon_{AC}\varepsilon_{BD}\varepsilon_{A'D'}\varepsilon_{B'C'} - i\varepsilon_{AD}\varepsilon_{BC}\varepsilon_{A'C'}\varepsilon_{B'D'} \tag{2.32}$$

for the Levi–Cività tensor. It is not hard to show that, conversely, *any* kinematic structure satisfying (2.29) and (2.31) is obtainable in the form (2.27) and that the freedom in Z^a is, in fact, precisely that given by (2.28).

The (restricted-Poincaré-invariant, pseudoscalar) quantity s of (2.31), called the *helicity* of the massless particle, turns out to be simply one-half of the twistor norm (2.8):

$$s = \tfrac{1}{2} Z^a \bar{Z}_a \tag{2.33}$$

Thus we see that right-handed ($s > 0$) massless particles are described by right-handed twistors and left-handed ($s < 0$) massless particles by left-handed twistors. When the helicity vanishes, the twistor is null, and in this case defines a unique light ray along which $\omega^A(x)$ vanishes. Referring back to (2.27), we see that these are the points x about which the angular momentum of the particle vanishes, and we can identify this light ray as the *world-line of the particle.* But when $s \neq 0$, the field is nowhere vanishing (for real x) and, furthermore, it turns out that every point of $\mathbb{M}$ is on an equal footing with every other, in the sense that there is a (restricted) conformal transformation taking one into the other while leaving the field $\omega^A(x)$ invariant. Thus when $s \neq 0$ the particle is not, in any conformally invariant sense, localizable. However, if a particular time axis is selected, with future-pointing unit timelike tangent vector t^a, this can be used to construct a t^a-dependent "norm"

$$t_{AA'}\omega^A(x)\bar{\omega}^{A'}(x) > 0 \tag{2.34}$$

the minimum values of which do, in fact, define a light ray that may be regarded as the approximate world-line of the particle [in fact, the same light

ray as that defined by the null twistor $(\omega^A - st^{AA'}\pi_{A'}(p_b t^b)^{-1}, \pi_{A'})$, the minimum value of (2.34) being $\frac{1}{2}s^2(p_b t^b)^{-1}$].

The kinematic structure of a particle (or system), whether massless or not, may be collected together in the form of the *kinematic twistor*

$$A^{\rho\sigma} = \begin{bmatrix} -2i\mu^{RS} & p^R{}_{S'} \\ p_{R'}{}^S & 0 \end{bmatrix} \tag{2.35}$$

where

$$M^{AA'} = \mu^{AB}\varepsilon^{A'B'} + \varepsilon^{AB}\bar{\mu}^{A'B'}, \qquad \mu^{AB} = \mu^{BA} \tag{2.36}$$

Every kinematic twistor is subject to

$$A^{\rho\sigma} = A^{\sigma\rho}, \qquad A^{\rho\tau}I_{\sigma\tau} = \bar{A}_{\sigma\tau}I^{\rho\tau} \tag{2.37}$$

In the particular case of a massless particle, the relations (2.37) take the form

$$A^{\rho\sigma} = 2Z^{(\rho}I^{\sigma)\tau}\bar{Z}_\tau \tag{2.38}$$

Note that since the relation (2.38) is not linear in the twistor Z^α, the (complex) linear structure of $\mathbf{T}$ is not directly reflected in any additive property of the kinematics. [Indeed, the sum of different expressions (2.38) will generally yield the kinematics of a massive system not a massless one.] In order to see the physical role of the linear structure of $\mathbf{T}$ it appears to be necessary to pass to yet one further interpretation of a twistor, namely, as a helicity-raising operator for massless fields. This is described in the next section.

2.2. Massless Free Fields

Let $\phi_{AB\cdots L}$ and $\psi_{A'B'\cdots L'}$ be symmetric spinor fields, each with n indices ($n = 1, 2, 3, \ldots$), being functions of the space–time point x. The massless free-field equations for spin $\frac{1}{2}n\hbar$ are

$$\nabla^{AA'}\phi_{AB\cdots L} = 0 \tag{2.39}$$

and

$$\nabla^{AA'}\psi_{A'B'\cdots L'} = 0 \tag{2.40}$$

If the fields are of positive frequency [which in the usual descriptions means composed of Fourier components of the form $\exp(-in_a x^a)$ with n_a future pointing) then the solutions of (2.39) describe wave functions for *negative helicity*, or *left-handed particles* ($s = -\frac{1}{2}n\hbar$), while the solutions of (2.40) refer to *positive helicity* or *right-handed particles* ($s = \frac{1}{2}n\hbar$). In any case, whether or not the positive-frequency condition holds, the solutions of (2.39) are referred to as *anti-self-dual fields* and the solutions of (2.40) as *self-dual fields*. The reason for this terminology arises from the fact that in

the case $n = 1$ we can define the bivector $F_{ab}(= -F_{ba}$, real if $\bar{\phi}_{A'B'} = \psi_{A'B'})$ by

$$F_{AA'BB'} = \phi_{AB}\varepsilon_{A'B'} + \varepsilon_{AB}\psi_{A'B'} \tag{2.41}$$

and find that its decomposition into anti-self-dual and self-dual parts,

$$F_{ab} = F_{ab}^{(-)} + F_{ab}^{(+)} \tag{2.42}$$

where

$$\begin{aligned} F_{ab}^{(\pm)*} &:= \tfrac{1}{2}F_{cd}^{(\pm)}e_{ab}{}^{cd} \\ &= \pm iF_{ab}^{(\pm)} \end{aligned} \tag{2.43}$$

is given by

$$F_{AA'BB'}^{(-)} = \phi_{AB}\varepsilon_{A'B'}, \qquad F_{AA'BB'}^{(+)} = \varepsilon_{AB}\psi_{A'B'} \tag{2.44}$$

The equations (2.39) and (2.40) together are, now, Maxwell's free-space equations[7]

$$\nabla_{[a}F_{bc]} = 0, \qquad \nabla^a F_{ab} = 0 \tag{2.45}$$

Similarly, in the case $n = 2$, the fields ϕ_{ABCD}, $\psi_{A'B'C'D'}$ define a tensor K_{abcd}, trace free with Riemann tensor symmetries, by

$$K_{AA'BB'CC'DD'} = \phi_{ABCD}\varepsilon_{A'B'}\varepsilon_{C'D'} + \varepsilon_{AB}\varepsilon_{CD}\psi_{A'B'C'D'} \tag{2.46}$$

Again, the first and second terms on the right define, respectively, the anti-self-dual and self-dual parts $K_{abcd}^{(\pm)}$ of K_{abcd}, where

$$\begin{aligned} K_{abcd}^{(\pm)*} &:= \tfrac{1}{2}K_{abpq}^{(\pm)}e_{cd}{}^{pq} = \tfrac{1}{2}K_{pqcd}^{(\pm)}e_{ab}{}^{pq} \\ &= \pm iK_{abcd}^{(\pm)} \end{aligned} \tag{2.47}$$

Now the equations (2.39) and (2.40) become the "Bianchi identity" spin-2 (linearized Einstein) field equations

$$\nabla_{[a}K_{bc]pq} = 0 \tag{2.48}$$

or, by virtue of the trace-free and symmetry conditions, equivalently,

$$\nabla^a K_{abcd} = 0 \tag{2.49}$$

For completeness we include also the zero-spin case, where the free-field equation is now the wave equation

$$\nabla_a\nabla^a\chi = 0 \tag{2.50}$$

We wish to examine the relation between the massless free-field equations (2.39), (2.40), (2.50) and the twistor equation (2.15). Let

$$\begin{aligned} Q^\alpha &= (\kappa^A(x), \tau_{A'}(x)) \\ &= (\kappa^A - ix^{AA'}\tau_{A'}, \tau_{A'}) \end{aligned}$$

be a twistor of valence $\begin{bmatrix}1\\0\end{bmatrix}$, using the description (2.19), so that $\kappa^A(x)$ satisfies the twistor equation

$$\nabla^{(A}_{A'}\kappa^{B)}(x) = 0$$

and

$$\nabla_{AA'}\kappa^B(x) = -i\tau_{A'}(x)\varepsilon_A{}^B \tag{2.51}$$

Now suppose that $\phi_{AB\cdots L}$ (symmetric) satisfies (2.39) for helicity $-\frac{1}{2}n\hbar$. Then assuming $n > 1$, the field

$$\chi_{B\cdots L} = \kappa^A(x)\phi_{AB\cdots L} \tag{2.52}$$

satisfies (2.39), for helicity $-\frac{1}{2}(n-1)\hbar$, precisely because the twistor equation holds! If $n = 1$, then χ satisfies (2.50), for the same reason. Now suppose that $\psi_{A'B'\cdots L'}$ (symmetric) satisfies (2.40), for helicity $\frac{1}{2}n\hbar$, and define

$$\theta_{A'B'\cdots L'M'} = -i\kappa^M(x)\nabla_{MM'}\psi_{A'B'\cdots L'} - (n+1)\psi_{(A'B'\cdots L'}\tau_{M')} \tag{2.53}$$

Then we find that $\theta_{A'\cdots M'}$ satisfies (2.40) for helicity $\frac{1}{2}(n+1)\hbar$, because of (2.51). Likewise, if χ satisfies (2.50), then

$$\theta_{A'} = -i\kappa^A(x)\nabla_{AA'}\chi - \chi\tau_{A'} \tag{2.54}$$

satisfies (2.40) for helicity $\frac{1}{2}\hbar$. Thus, in each case, the twistor Q^α provides a way of *raising the helicity by* $\frac{1}{2}\hbar$ (the positive-frequency property of the fields being undisturbed by these operations), so we may interpret a valence $\begin{bmatrix}1\\0\end{bmatrix}$ twistor as a helicity-raising operator. Similarly, by taking complex conjugates, we obtain an interpretation of a valence $\begin{bmatrix}0\\1\end{bmatrix}$ twistor as a helicity-lowering operator, which reduces the helicity by $\frac{1}{2}\hbar$. Notice that in each case the operations are *linear* in the twistor, so the complex linear structure of $\mathbb{T}$ and of $\mathbb{T}^*$ can be directly interpreted. And since we are primarily concerned with wave functions here, we see that this linear structure is essentially bound up with the *complex linearity of quantum superposition.*

These twistor operations are closely associated with a basic twistor construction whereby solutions of (2.39), (2.40), and (2.50) may be written down by means of a contour integral expression. Let $f(Z^\alpha)$ be a holomorphic function of a twistor Z^α of homogeneity degree either $n-2$ or $-n-2$, where n is some non-negative integer:

$$f(\lambda Z^\alpha) = \lambda^{n-2}f(Z^\alpha), \qquad \text{or} \qquad f(\lambda Z^\alpha) = \lambda^{-n-2}f(Z^\alpha) \tag{2.55}$$

The precise domain $\mathcal{D} \subset \mathbb{T}$ over which these functions are defined will, for the moment, be left unspecified, except that it is to be adequate for the

following contour integrals† to be nontrivially defined:

$$\phi_{AB\cdots L} = \frac{1}{(2\pi i)^2}\oint \frac{\partial}{\partial\omega^A}\frac{\partial}{\partial\omega^B}\cdots\frac{\partial}{\partial\omega^L} f(Z^\alpha)\, d^2\pi \tag{2.56}$$

or

$$\psi_{A'B'\cdots L'} = \frac{1}{(2\pi i)^2}\oint \pi_{A'}\pi_{B'}\cdots\pi_{L'} f(Z^\alpha)\, d^2\pi \tag{2.57}$$

where

$$Z^\alpha = (\omega^A, \pi_{A'}) = (ix^{AA'}\pi_{A'}, \pi_{A'}) \tag{2.58}$$

(origin held fixed at O) and

$$d^2\pi = d\pi_{0'}\wedge d\pi_{1'} = \tfrac{1}{2}\varepsilon^{A'B'}\, d\pi_{A'}\wedge d\pi_{B'} \tag{2.59}$$

Thus there is to be some nonempty open set $\mathscr{E}$ in $\mathbb{CM}^I$ on which the field is to be defined, where for each $x \in \mathscr{E}$ there is a nontrivial two-real-dimensional closed surface (contour) in $\mathscr{D}$ consisting of twistors Z^α incident with x. For each fixed x, we may regard f as a function of $\pi_{A'}$, i.e., a function on some open subset $\mathscr{F}_x$ of $\tilde{\mathbb{S}}^*$, and the contour is to be taken in $\mathscr{F}_x$ and to vary continuously with x. For a nontrivial field it is necessary that the homology class of the contour be nonzero. For by the holomorphic property of f, (2.56) and (2.57) are genuine contour integrals, i.e., invariant under continuous deformations of the contour. It is now a trivial matter to verify that the field equations (2.39) and (2.40) are automatically satisfied in each case—as also is (2.50), allowing $n = 0$—because the operator $\nabla_{AA'}$ becomes, under the integral sign

$$\nabla_{AA'} \rightsquigarrow i\pi_{A'}\frac{\partial}{\partial\omega^A} \tag{2.60}$$

Let $Q^\alpha = (\kappa^A(x), \tau_{A'}(x))$ be a twistor of valence $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$. Then the linear map

$$f(Z^\alpha) \mapsto Q^\beta \frac{\partial}{\partial Z^\beta} f(Z^\alpha)$$

is a helicity-raising operator: when combined with the contour integral formulas (2.56) and (2.57), it yields the helicity-raising operations (2.52) and (2.53), respectively. Similarly, a twistor R_α of valence $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$ determines a helicity-lowering operation via

$$f(Z^\alpha) \mapsto R_\beta Z^\beta f(Z^\alpha)$$

† The slightly more familiar (one-dimensional) form of these integrals is given in (2.63), (2.64). The form introduced here has advantages in certain contexts, however (cf. References 2 and 9). The introduction of the $\partial/\partial\omega$ terms, leading to the expression (2.56), followed a suggestion made by L. P. Hughston.

With the given homogeneities (2.55), we can also write down a slightly simplified version of (2.56) and (2.57), when the two-real-dimensional contour is replaced, for each x^a, by a one-dimensional closed loop [and, in fact, with a wrong value of homogeneity (2.55), the result of the integral would necessarily be zero, having to be independent of $\pi_{A'}$]. Choose any portion of the integration surface at which $\pi_{0'} \neq 0$ and rewrite the differential in (2.59):

$$\frac{1}{2\pi i} d^2\pi = (\pi_{0'})^2 \left(\frac{1}{2\pi i} \frac{d\pi_{0'}}{\pi_{0'}}\right) \wedge d\left(\frac{\pi_{1'}}{\pi_{0'}}\right) \tag{2.61}$$

Owing to the -2 homogeneity of each integrand in (2.56) and (2.57) (excluding the differential $d^2\pi$ itself), we find that $(\pi_{0'})^2$ times this integrand is a function of the one variable $\pi_{1'}/\pi_{0'}$. Thus, assuming that the contour can be deformed into one that, for each fixed value of $\pi_{1'} : \pi_{0'}$, is simply a closed loop about the origin, the $\pi_{0'}$ dependence can be integrated out explicitly, using Cauchy's formula, which simply eliminates the factor $(2\pi i)^{-1}(d\pi_{0'})/\pi_{0'}$. Thus the differential (2.61) gets replaced by

$$\begin{aligned}(\pi_{0'})^2\, d\left(\frac{\pi_{1'}}{\pi_{0'}}\right) &= \pi_{0'}\, d\pi_{1'} - \pi_{1'}\, d\pi_{0'} \\ &= \varepsilon^{A'B'} \pi_{A'}\, d\pi_{B'} \\ &=: \Delta\pi \end{aligned} \tag{2.62}$$

The corresponding result also holds with $(\pi_{0'}, \pi_{1'})$ replaced by $(\pi_{1'}, -\pi_{0'})$, so we can also cover regions where $\pi_{0'} = 0$. The upshot is that we can rewrite (2.56) and (2.57), respectively, as

$$\phi_{A\cdots L} = \frac{1}{2\pi i} \oint \frac{\partial}{\partial \omega^A} \cdots \frac{\partial}{\partial \omega^L} f(Z^a) \Delta\pi \tag{2.63}$$

and

$$\psi_{A'\cdots L'} = \frac{1}{2\pi i} \oint \pi_{A'} \cdots \pi_{L'} f(Z^a) \Delta\pi \tag{2.64}$$

The integration may now be regarded as taking place in $\mathbb{PT}$, rather than $\mathbb{T}$. Each x defines a projective line, having topology S^2, in $\mathbb{PT}$ and the contour is a closed loop on this S^2 separating singularities of f.

In order to be somewhat more explicit about the domain and the location of the contour, let us take the particular case when the fields are of positive frequency. One way of stating this positive-frequency condition is to assert[8] that the domain of definition of the field extends holomorphically to the entire *forward tube* $\mathbb{CM}^+$, whose points are those elements of $\mathbb{CM}^I$ with position vectors of the form

$$z^a = x^a - iy^a \tag{2.65}$$

where x^a and y^a are real, with y^a timelike and future pointing:

$$y_a y^a > 0, \qquad y^0 > 0 \tag{2.66}$$

The extended field is to be holomorphic on $\overline{\mathbb{CM}^+}$ (the closure of $\mathbb{CM}^+$ in $\mathbb{CM}$), for an *analytic* positive-frequency field, or merely holomorphic on $\mathbb{CM}^+$ if the field need not be analytic but only C^∞, or C^1, say, or a distribution or even a hyperfunction.[24] Observe that each point $z \in \mathbb{CM}^+$ is represented in $\mathbb{PT}$ by a line $\mathbf{z}$ which lies entirely in $\mathbb{PT}^+$ (and conversely). This follows from the fact that

$$\begin{aligned} Z^a \bar{Z}_a &= i(x^{AA'} - iy^{AA'})\pi_{A'}\bar{\pi}_A - i\pi_{A'}(x^{AA'} + iy^{AA'})\bar{\pi}_A \\ &= 2y^a p_a > 0 \end{aligned} \tag{2.67}$$

whenever $Z^\alpha = (iz^{AA'}\pi_{A'}, \pi_{A'})$.

Suppose that $\overline{\mathbb{PT}^+}$ is covered by two open sets $\mathcal{U}_1, \mathcal{U}_2 \subset \mathbb{PT}$, so

$$\overline{\mathbb{PT}^+} \subset \mathcal{U}_1 \cup \mathcal{U}_2 \tag{2.68}$$

with the property that every line $\mathbf{z}$ in $\overline{\mathbb{PT}^+}$ meets their intersection

$$\mathcal{D}_{12} = \mathcal{U}_1 \cap \mathcal{U}_2 \tag{2.69}$$

in an annular region. For example, we can choose

$$\mathcal{U}_1 = \{Z^\alpha \in \mathbb{T} |\ (1+\varepsilon)|Z^0 + Z^2| > |Z^1 + Z^3|,\ \varepsilon|Z^0 + Z^1|^2 + Z^\alpha \bar{Z}_\alpha > 0\} \tag{2.70}$$

$$\mathcal{U}_2 = \{Z^\alpha \in \mathbb{T} |\ |Z^0 + Z^2| < (1+\varepsilon)|Z^1 + Z^3|,\ \varepsilon|Z^1 + Z^3|^2 + Z^\alpha \bar{Z}_\alpha > 0\}$$

where ε is some positive number (see Figure 4). Then, if f is holomorphic in $\mathcal{D}_{12}$, we can perform the required contour integral in (2.63) and (2.64), x now being replaced by the *complex* point z, by choosing a closed loop ($\cong S^1$)

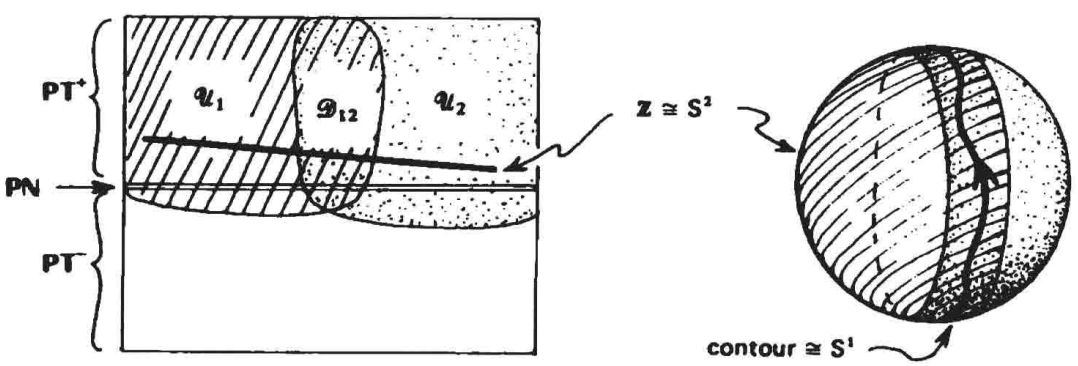


Fig. 4

which goes once around the annular region $\mathcal{D}_{12} \cap \mathbf{z}$, with an orientation defined by the ordering $\mathcal{U}_1$, $\mathcal{U}_2$. It can be shown, in fact, that *any* analytic positive-frequency field (at least for $s > 0$) on $\mathbb{M}$ can be obtained in this way, where $\mathcal{U}_1$ and $\mathcal{U}_2$ are the *specific* open sets (2.70), but where the value of ε may depend on the field.

One special class of fields that can be readily obtained by this construction are the *elementary states*, for which f has the form

$$f(Z^\alpha) = \frac{(A_\alpha Z^\alpha)^a (B_\beta Z^\beta)^b}{(C_\gamma Z^\gamma)^{c+1} (D_\delta Z^\delta)^{d+1}} \tag{2.71}$$

with $c + d - a - b = 2s\hbar^{-1}$, where a, b, c, d are non-negative integers, the twistors C_α and D_α defining planes in $\mathbb{PT}$ whose intersection lies entirely in $\mathbb{PT}^-$ (see Figure 5), and where $A_\alpha, B_\alpha, C_\alpha, D_\alpha$ are linearly independent. (In previous work[9] the term "elementary state" has been reserved for the case $a = b = 0$, but the restriction is dropped, here, in order to include the negative helicity fields in a systematic way.) The sets $\mathcal{U}_1$ and $\mathcal{U}_2$ are the portions of $\mathbb{PT}^+$ not lying on the planes **D** and **C**, respectively. In fact, for *fixed* $A_\alpha, B_\alpha, C_\alpha, D_\alpha$ and for each fixed integer $2s\hbar^{-1}$, the fields obtained by inserting (2.71) into (2.63), (2.64) provide a *basis* for positive-frequency massless fields (wave functions) of helicity s.

Another special class of fields is obtained by selecting a complex-analytic surface $\mathcal{Q}$ in $\overline{\mathbb{PT}^+}$ (say), defined, locally, by some holomorphic homogeneous equation

$$q(Z^\alpha) = 0, \qquad Z^\alpha \in \overline{\mathbb{T}^+} \tag{2.72}$$

where the zeros of q in $\overline{\mathbb{T}^+}$ are simple. We suppose that $\mathcal{Q}$ meets every line in $\overline{\mathbb{PT}^+}$ exactly once. (Many such surfaces $\mathcal{Q}$ do in fact exist.) Then put

$$g(Z^\alpha) = p(Z^\alpha)\{q(Z^\alpha)\}^{-r} \tag{2.73}$$

(for some positive integer r), where p and q are both holomorphic in some

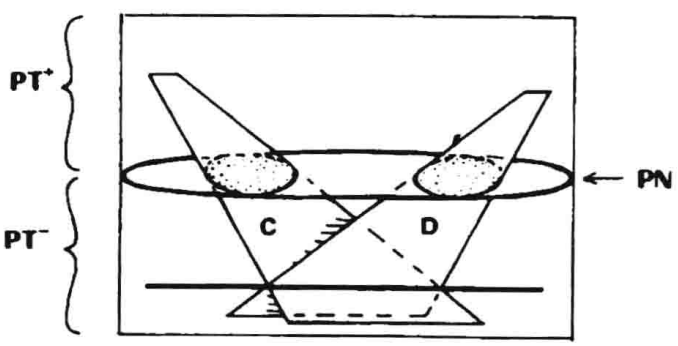


Fig. 5

neighborhood $\mathcal{U}_1$ of $\mathcal{Q}$ and where $\mathcal{U}_2 \cap \mathbb{PT}^+ = \mathbb{PT}^+ - \mathcal{Q}$. When $s > 0$ and $r = 1 < n$, the fields obtained by inserting (2.73) into (2.64) have a special interest in that they are *null*.[10] (In fact, non-positive-frequency null fields defined on other portions of $\mathbb{CM}$ can be given by an exactly corresponding construction but where where the special relation to $\mathbb{PT}^+$ is dropped.) A null field $\psi_{A'\cdots L'}$ (with $n > 1$ indices) is defined to be one having the form

$$\psi_{A'B'\cdots L'} = \lambda \alpha_{A'} \alpha_{B'} \cdots \alpha_{L'} \tag{2.74}$$

at every point of $\mathbb{CM}$ at which the field is defined. Another way of expressing this condition is

$$\alpha^{A'} \psi_{A'B'\cdots L'} = 0 \tag{2.75}$$

for some $\alpha^{A'} \neq 0$. We observe that, with (2.73) inserted into (2.64), the expression in (2.75) indeed vanishes provided that $\alpha_{A'}$ is proportional to the value of $\pi_{A'}$ obtained at the intersection point $\mathbf{z} \cap \mathcal{Q}$. If $r > 1$, the same holds if (2.75) is replaced by

$$\alpha^{A'} \cdots \alpha^{C'} \psi_{A'\cdots C'D'\cdots L'} = 0 \tag{2.76}$$

where there are r α-factors. (This is the condition that $\psi_{\ldots}$ have at least $n - r + 1$ coincident principal null directions[10] at the point $z \in \mathbb{M}^I$. If $r > n$, such a field is called algebraically special.)

The set $\mathcal{Q} \cap \mathbb{PN}$ is three-real-dimensional and defines a congruence of light rays in $\mathbb{M}$. The Robinson congruence (2.12) is a special case, where here $q(Z^\alpha) \equiv W_\alpha Z^\alpha$. A general property of all such congruences (defined by a holomorphic $\mathcal{Q}$) is that they are *shear-free*; the converse result[5] is that such a holomorphic $\mathcal{Q}$ always exists, locally, for any *analytic* shear-free congruence of light rays (a theorem of Kerr). It may be remarked that the shear-free condition always holds for the $\alpha^{A'}$ congruence, by virtue of (2.76) and (2.40), provided that $r < n$.

A striking feature of the twistor description of massless fields, according to (2.56), (2.57) or (2.63), (2.64) is that the field equations (2.39), (2.40) are automatically satisfied, the twistor function f having merely to possess the homogeneity property (2.55). The twistor function thus gives a very simple way of representing and generating solutions of the field equations. This virtue is, to some extent, offset by the rather strange relation that the domain $\mathcal{D}(=\mathcal{D}_{12})$ has to the domain on which the space–time field is defined, to the evident nonuniqueness of this domain [for example, any $SU(2, 2)$ action on the region $\mathcal{U}_1 \cap \mathcal{U}_2$ of (2.70) would clearly yield another domain of equal validity] and to the nonuniqueness of the twistor function even when the domain is held fixed. Let us examine this nonuniqueness for a twistor function f. We first note that if f were in fact holomorphic over the whole of $\mathcal{U}_1$ then the result of the integration (2.63) or (2.64) would vanish because (cf. Figure 4) the contour could be contracted to a point within

the holomorphic region. The same would apply if f were holomorphic over $\mathcal{U}_2$. Hence, if we make the replacement

$$f \mapsto f + h_1 - h_2 \tag{2.77}$$

where h_i is holomorphic on $\mathcal{U}_i (i = 1, 2)$, f being holomorphic on $\mathcal{U}_1 \cap \mathcal{U}_2$, then the same field (2.63), (2.64) will result as before.

We can generalize this situation. Suppose we are concerned with some region $\mathcal{R}$ in $\mathbb{T}$, such as $\mathbb{T}^+$ or $\overline{\mathbb{T}}^+$, which is swept out by the linear 2-spaces representing the points of some appropriate region in $\mathbb{CM}$. Suppose that $\mathcal{R}$ is covered by open sets $\mathcal{U}_1, \mathcal{U}_2, \mathcal{U}_3, \ldots$, possibly infinite in number, but, if so, locally finite (i.e., with only a finite number of the $\mathcal{U}_i$ containing any given point). Now instead of having a single holomorphic twistor function we envisage such a function f_{ij} for every pair of sets $\mathcal{U}_i$, $\mathcal{U}_j$ [each subject to the same homogeneity restriction (2.55)] where the domain of f_{ij} is $\mathcal{U}_i \cap \mathcal{U}_j$ and where

$$f_{ij} = -f_{ji} \qquad (i, j = 1, 2, \ldots) \tag{2.78}$$

Using the symbol ρ_i to denote *restriction* of the domain by intersection with $\mathcal{U}_i$,† we require the *cocycle condition*

$$\rho_{[i} f_{jk]} = 0 \qquad (i, j, k = 1, 2, \ldots) \tag{2.79}$$

to hold (which written out in full is $\rho_i f_{jk} + \rho_j f_{ki} + \rho_k f_{ij} = 0$). Such a set of functions $\{f_{ij}\}$ is called a 1-*cocycle* for the covering $\{\mathcal{U}_i\}$ of $\mathcal{R}$, with coefficients in the sheaf $\mathcal{O}$ (of germs of holomorphic functions on $\mathbb{T}$). It is called a 1-cochain if merely the skew symmetry (2.78) is required to hold. A cocycle is called a 1-*coboundary* for this covering if it has the form

$$f_{ij} = \rho_{[i} h_{j]} \tag{2.80}$$

[and we observe that (2.79) follows automatically from this because $\rho_{[i}\rho_{j]} = 0$], the 0-cochain $\{h_i\}$ being a system of holomorphic functions h_i with domain $\mathcal{U}_i$. We define the first cohomology group

$$H^1_{\{\mathcal{U}_i\}}(\mathcal{R}, \mathcal{O}) \tag{2.81}$$

for the covering $\{\mathcal{U}_i\}$ of $\mathcal{R}$, with coefficients in $\mathcal{O}$, to be the additive group (or vector space over $\mathbb{C}$) of 1-cocycles *modulo* 1-coboundaries, i.e., we regard two families $\{f_{ij}\}$, $\{\tilde{f}_{ij}\}$ as defining equivalent cohomology groups elements if $\tilde{f}_{ij} - f_{ij} = \rho_{[i} h_{j]}$ for some family $\{h_i\}$.

Note that the replacement (2.77) simply passes from one such cochain to an equivalent one, where there are here just *two* $\mathcal{U}_i$ and $f = f_{12}$ (the cocycle condition being now vacuous). In fact, our contour integral formulas can be generalized to the case where there are more than two $\mathcal{U}_i$, the contours being now generalized to *branched contours* as indicated in Figure

† This notation is due to L. P. Hughston.

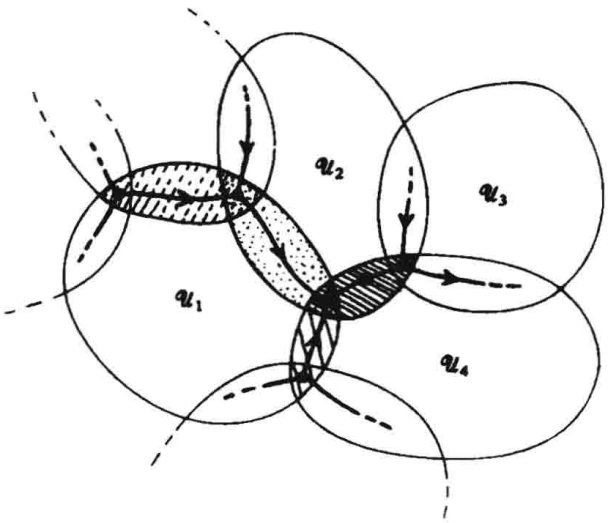


Fig. 6

6. We shall not go into this in detail here, but give another method for evaluating the space–time field from the cocycle in a moment. The point is that the contributions from 1-coboundaries to these fields is always zero, so the field is really a function of an element of the cohomology group (2.81) rather than of the cocycle. This, however, still depends on the particular covering $\{\mathcal{U}_i\}$, in general, and not simply on $\mathcal{R}$. The dependence on the covering is removed once we pass to the *direct limit* for finer and finer coverings, to obtain

$$H^1(\mathcal{R}, \mathcal{O}) \tag{2.82}$$

the *first cohomology group of* $\mathcal{R}$ *with coefficients in* $\mathcal{O}$. Fortunately it is often not necessary to take this limit, this being the case if $\{\mathcal{U}_i\}$ is, in a suitable sense, "fine enough." In the present context, this will be the case if each $\mathcal{U}_i$ is what is called a *Stein manifold* or *domain of holomorphy*.[11] In any case, the dependence on the covering is now removed, so we have an invariant way of talking about a "twistor function" as an element of (2.82). When $\mathcal{R} = \mathbb{T}^+$, this gives us precisely the wave functions in $\mathbb{CM}^+$.

We do not go into these questions in detail here, but end this section by giving a "splitting method" whereby the required space–time field may be obtained from any representative cochain $\{f_{ij}\}$. This procedure is based on the fact that

$$H^1(\mathbb{C}^2 - \{0\}, \mathcal{O}(m)) = \begin{cases} 0 & \text{if } m \geqslant -1 \\ \mathbb{C}^{-m-1} & \text{if } m < -1 \end{cases} \tag{2.83}$$

where $\mathcal{O}(m)$ is the sheaf of germs of holomorphic functions on $\mathbb{C}^2 - \{0\}$,

homogeneous of degree m in the two complex variables. So in particular,

$$H^1(\mathbb{C}^2 - \{0\}, \mathcal{O}(-1)) = 0 \tag{2.84}$$

which means that every 1-cocycle f_{ij} homogeneous of degree -1 on $\mathbb{C}^2 - \{0\}$ is a 1-coboundary: f_{ij} has the form

$$f_{ij} = \rho_{[i}h_{j]}$$

for some 1-cochain h_j.

Suppose now that $\{f_{ij}(Z^\alpha)\}$ is a 1-cocycle homogeneous of degree $-n-2$ ($n \geq 0$) in Z^α. We want to use f_{ij} to define a massless field $\psi_{A'\cdots L'}(x^a)$. (The procedure for homogeneity $n-2$ cocycles is similar and will not be entered into here.) Let the twistor Z^α be incident with the space–time point x^a (i.e., put $\omega^A = ix^{AA'}\pi_{A'}$) and multiply f_{ij} by $(n+1)$ $\pi_{A'}$ spinors: this yields

$$\pi_{A'}\cdots\pi_{C'}f_{ij}(ix^{DD'}\pi_{D'}, \pi_{D'}) \tag{2.85}$$

Now for *fixed* $x^{AA'}$, the expression (2.85) represents a 1-cocycle on $\hat{\mathbb{S}}^* - \{0\} = \mathbb{C}^2 - \{0\}$, homogeneous of degree -1. In view of (2.84) it is also a 1-coboundary; in other words there exist functions $h_{jA'\cdots C'}(x^a, \pi_{D'})$ such that

$$\pi_{A'}\cdots\pi_{C'}f_{ij} = \rho_{[i}h_{j]A'\cdots C'} \tag{2.86}$$

Define $\psi_{A'\cdots C'}$ by

$$\psi_{A'\cdots C'}(x^e) = \pi^{D'}h_{jA'\cdots C'D'}(x^e, \pi_{E'}) \tag{2.87}$$

Although the right-hand side of (2.87) appears to depend on the suffix j and on $\pi_{E'}$, it is, by virtue of (2.86), independent of j and $\pi_{E'}$. Furthermore, the field $\psi_{A'\cdots C'}$ defined in this way is a solution of the massless-field equations 2.40). This follows because $H^0(\mathbb{C}^2 - \{0\}, \mathcal{O}(-1)) = 0$ [compare (2.84)]. In the special case when the covering of twistor space consists of only two patches $\mathcal{U}_1$ and $\mathcal{U}_2$, and $f = f_{12}$, the procedure just described yields the same field $\psi_{A'\cdots C'}$ as the contour integral formula (2.57).

3. Twistors in Curved Space–Time

In the previous section we saw how the complex linear structure of $\mathbb{PT}$ corresponds to the conformal structure of flat space–time and how holomorphic twistor functions correspond to certain representations of the conformal group $C(1, 3)$ (namely, massless free fields). Can twistor theory be generalized to curved space–times in such a way that these features are preserved? At first sight, there is no reason for believing that it can, because the conformal group does not act on curved space–time: a curved space–time in general has *no* conformal symmetries. So one certainly has to forget about the group theory/linear structure aspects and concentrate instead on the *complex structure* of twistor space. Even then one runs into obstacles, as

 Twistors for flat and curved space-time. *General relativity and gravitation, Vol. 2,* pp. 283–328, Plenum, New York and London, 1980

we shall see below. Nevertheless, there have been some remarkable partial successes and these form the subject of Sections 4 and 5 of this chapter.

To investigate the problem, let us consider the space $\mathbb{P}\mathcal{N}$ of unscaled null geodesics in a space–time M. In order to avoid possible pathologies (for example, the space $\mathbb{P}\mathcal{N}$ may in general be non-Hausdorff), assume that M is globally hyperbolic,[12] with a Cauchy surface diffeomorphic to $\mathbb{R}^3$. As a real manifold, $\mathbb{P}\mathcal{N}$ is then diffeomorphic to $\mathbb{R}^3 \times S^2$ (since through each point of the Cauchy surface there passes a 2-sphere's worth of null geodesics). What additional structure can we identify on $\mathbb{P}\mathcal{N}$? To motivate our search, consider the Sachs equations[13] which relate to a congruence of null geodesics:

$$\begin{aligned} D\rho &= \rho^2 + \sigma\bar{\sigma} + \Phi \\ D\sigma &= (\rho + \bar{\rho})\sigma + \Psi \end{aligned} \tag{3.1}$$

Here D means differentiation along the geodesics in the congruence; Re (ρ) measures the divergence of congruence, Im (ρ) its rotation, and σ its shear; Φ is a (real) component of the Ricci tensor, and Ψ is a (complex) component of the Weyl conformal curvature tensor. What we are looking for are properties or quantities essentially preserved by D, i.e., preserved as one moves along the null geodesics. For this would correspond to intrinsic structure in $\mathbb{P}\mathcal{N}$. The most obvious such quantity is Im (ρ), the *rotation* (or twist) of the congruence. So we expect to find some "invariant structure" on $\mathbb{P}\mathcal{N}$ which measures rotation. This structure may be described as follows.

It is in fact more convenient to deal with the space $\mathcal{N}$ of *scaled* null geodesics in M. The canonical symplectic form $dp_a \wedge dx^a$ on the cotangent bundle of M is constant along the lifts of null geodesics (Lagrange's identity[14]) and so it projects down to a 2-form ω on $\mathcal{N}$. This 2-form is the rotation-measuring structure on $\mathcal{N}$ that we are looking for. One way of seeing this is to observe that if $p_a = p_a(x)$ is the cotangent vector field of a congruence of null geodesics, then we have

$$dp_a(x) \wedge dx^a = \nabla_{[a}p_{b]}\, dx^a \wedge dx^b$$

and $\nabla_{[a}p_{b]}$ is the rotation (or twist) of p_a.

In flat space–time, the twistor version of this structure is easily found. The twistor coordinates $Z^\alpha \leftrightarrow (\omega^A, \pi_{A'})$ are related to p_a and x^a by

$$\begin{aligned} p_a &= \bar{\pi}_A \pi_{A'} \\ \omega^a &= ix^{AA'}\pi_{A'} \qquad \text{with } x^{AA'} = \bar{x}^{AA'} \end{aligned}$$

Thus

$$\begin{aligned} i\bar{Z}_\alpha\, dZ^\alpha &= -\bar{\pi}_A(dx^{AA'}\pi_{A'} + x^{AA'}\, d\pi_{A'}) + x^{AA'}\bar{\pi}_A\, d\pi_{A'} \\ &= -p_a\, dx^a \end{aligned}$$

and

$$dp_a \wedge dx^a = i\, dZ^\alpha \wedge d\bar{Z}_\alpha$$

The space $\mathcal{N}$ is obtained from the space of nonprojective null twistors by factoring out the phase circles

$$Z^\alpha \mapsto e^{i\theta} Z^\alpha \qquad \theta \text{ real}$$

and $\omega = i\, dZ^\alpha \wedge d\bar{Z}_\alpha$ is a symplectic form on $\mathcal{N}$.

Can we identify any other invariant structure on $\mathcal{N}$ or $\mathbb{P}\mathcal{N}$? In particular we might look for a *CR (Cauchy–Riemann) structure.* One can think of a CR structure on $\mathbb{P}\mathcal{N}$ as that structure which arises as a result of $\mathbb{P}\mathcal{N}$ being embedded in a complex manifold.[15] In the case of flat twistor space, of course, $\mathbb{P}\mathcal{N}$ is embedded in $\mathbb{PT}$; so $\mathbb{PN}$ has a CR structure and, as remarked in Section 2, this structure is closely related (by the Kerr theorem) to the condition that a null congruence be *shear-free.* However, equation (3.1) tells us that if $\Psi \neq 0$, then the shear-free condition $\sigma = 0$ is *not preserved* as one moves along the null geodesics. The shear-free condition can in general refer only to *one point* on each null geodesic, rather than to the geodesic in its entirety. This indicates that the space $\mathbb{P}\mathcal{N}$ does *not,* in general, admit a natural CR structure which relates to the space–time geometry via a "Kerr theorem." This is the basic obstacle to the generalization of twistor theory to curved space–time. Whether it can be overcome is not known. But there are at least two ways of sidestepping it:

(a) We could *complexify* the space–time and then impose the condition that, say, $\tilde{\Psi}_{A'B'C'D'} = 0$ while $\Psi_{ABCD} \neq 0$. This allows the existence of congruences with $\tilde{\sigma} = 0$ (but $\sigma \neq 0$). It turns out that the twistor formalism generalizes very neatly to such half-conformally-flat space–times, as we shall see in Section 4.

(b) On the other hand, we could select a point on each null geodesic, for example by choosing a spacelike hypersurface in the space–time. This leads to the theory of hypersurface twistors, which we shall discuss in Section 5.

It is worth mentioning that the shift in CR structure (or shear structure) can be seen very clearly in curved space–times describing *impulsive gravitational waves.*[7] These are constructed by taking two pieces M_1 and M_2 of Minkowski space–time and joining them across a null hypersurface $\mathcal{S}$. This is done in such a way that the metric is continuous across $\mathcal{S}$, but there is a δ function in the curvature. In other words, the impulsive gravitational wave resides on $\mathcal{S}$. Let Z^α denote twistors in the flat space M_1 and $\hat{Z}^\alpha$ twistors in M_2. By examining how null geodesics scatter through $\mathcal{S}$, one can derive a transformation law of the form

$$\begin{aligned} \hat{Z}^\alpha &= Z^\alpha - i\frac{\partial H}{\partial \bar{Z}_\alpha} \\ \hat{\bar{Z}}_\alpha &= \bar{Z}_\alpha + i\frac{\partial H}{\partial Z^\alpha} \end{aligned} \tag{3.2}$$

where $H = H(Z^\alpha, \bar{Z}_\alpha)$ is real valued, homogeneous of degree one in each of Z^α and $\bar{Z}_\alpha$, and satisfies $(\partial H/\partial Z^\alpha)(\partial H/\partial \bar{Z}_\alpha) = 0$. Clearly, (3.2) preserves the nullness condition $Z^\alpha \bar{Z}_\alpha = 0$ and the symplectic form $i\, dZ^\alpha \wedge d\bar{Z}_\alpha$; but it does *not* preserve the complex structure of twistor space.

An intriguing feature of this example is that Einstein's vacuum equations $R_{ab} = 0$ are satisfied if and only if H has the form

$$H(Z, \bar{Z}) = \bar{Z}_\alpha I^{\alpha\beta} \frac{\partial g}{\partial Z^\beta} + Z^\alpha I_{\alpha\beta} \frac{\partial \bar{g}}{\partial \bar{Z}_\beta}$$

where $g = g(Z^\alpha)$ is *holomorphic* and $I^{\alpha\beta}$, $I_{\alpha\beta}$ are the infinity twistors. Whether or not this indirect appearance of holomorphic functions is a general feature is not known.

4. The Nonlinear Gravitation Construction

4.1. Half-Flat Spaces and Curved Twistor Space

The curvature tensor of space–time may be decomposed into its irreducible (spinor) components as follows[7]:

$$\begin{aligned} R_{abcd} = {} & \Psi_{ABCD}\varepsilon_{A'B'}\varepsilon_{C'D'} + \bar{\Psi}_{A'B'C'D'}\varepsilon_{AB}\varepsilon_{CD} + \varepsilon_{AB}\Phi_{CDA'B'}\varepsilon_{C'D'} \\ & + \varepsilon_{A'B'}\Phi_{ABC'D'}\varepsilon_{CD} + 2\Lambda(\varepsilon_{AC}\varepsilon_{BD}\varepsilon_{A'C'}\varepsilon_{B'D'} - \varepsilon_{AD}\varepsilon_{BC}\varepsilon_{A'D'}\varepsilon_{B'C'}) \end{aligned} \tag{4.1}$$

where the curvature spinors Ψ_{ABCD}, $\Phi_{ABC'D'}$ and Λ satisfy

$$\begin{aligned} \Psi_{ABCD} &= \Psi_{(ABCD)} \\ \Phi_{ABC'D'} &= \bar{\Phi}_{ABC'D'} = \Phi_{(AB)(C'D')} \\ \Lambda &= \bar{\Lambda} \end{aligned} \tag{4.2}$$

Einstein's vacuum equations $R_{ab} = 0$ are equivalent to

$$\Phi_{ABC'D'} = 0 \qquad \Lambda = 0 \tag{4.3}$$

and Ψ_{ABCD} and $\bar{\Psi}_{A'B'C'D'}$ correspond, respectively, to the anti-self-dual and self-dual parts of the Weyl conformal curvature tensor:

$$C_{abcd} = C^-_{abcd} + C^+_{abcd} \tag{4.4}$$

where

$$\begin{aligned} C^-_{abcd} &= \Psi_{ABCD}\varepsilon_{A'B'}\varepsilon_{C'D'} \\ C^+_{abcd} &= \bar{\Psi}_{A'B'C'D'}\varepsilon_{AB}\varepsilon_{CD} \\ (C^\pm_{abcd})^* &:= \tfrac{1}{2}\varepsilon_{ab}{}^{ef}C_{efcd} = \pm iC^\pm_{abcd} \end{aligned} \tag{4.5}$$

Note that $\overline{C^+_{abcd}} = C^-_{abcd}$.

Consider now the curvature of a complex space–time M. In other words, M is a four-dimensional complex manifold with a complex, holomorphic, nonsingular metric g_{ab}. Since g_{ab} is complex one cannot assign a "signature" to it: the most one can require is that it be nonsingular, i.e., invertible. The complex space–time might or might not have a *Lorentz slice*, i.e., a four-real-dimensional submanifold of M in which g_{ab} is real and has Lorentz signature. If M *does* have a Lorentz slice $M_{\mathbb{R}}$, then we say that M is a *complexification* of $M_{\mathbb{R}}$. Any real analytic space–time can be complexified. One should think of the process of complexification as "thickening $M_{\mathbb{R}}$ into the complex." In general, if one tries to thicken out M too far, one encounters singularities.

Returning to the matter of the curvature tensor decomposition, the equations (4.1)–(4.5) carry over to a complex space, provided we make the following alterations:

(i) $\bar{\Psi}_{A'B'C'D'}$ is replaced by a new spinor field $\tilde{\Psi}_{A'B'C'D'}$, which is no longer the complex conjugate of Ψ_{ABCD} (so C^{+}_{abcd} and C^{-}_{abcd} are no longer conjugate).

(ii) The reality condition on Λ and hermiticity condition on $\Phi_{ABC'D'}$ are dropped.

A complex space-time is said to be *right-conformally flat* if $\tilde{\Psi}_{A'B'C'D'} = 0$ (i.e., if C_{abcd} is anti-self-dual). Clearly a right-conformally-flat space has a Lorentz slice if and only if it is *conformally flat*, so right-conformally-flat spaces are in this sense "essentially complex." If both $\tilde{\Psi}_{A'B'C'D'}$ and the Ricci tensor R_{ab} vanish, we say that the space is *right-flat*. Left-conformally flat and left-flat spaces are defined similarly to mean $\Psi_{ABCD} = 0$ and $\Psi_{ABCD} = 0 = R_{ab}$, respectively. Spaces of these types arose originally as the "$\mathscr{H}$ spaces" associated with asymptotically flat space–times.[3] Another possible application of right-flat spaces is that they may represent single-graviton helicity eigenstates[16]—for this reason, these spaces are sometimes referred to as "nonlinear gravitons."

The key to the generalization of twistor theory to half-conformally-flat spaces is to generalize the concept of a totally null 2-plane. A complex 2-surface Z in a complex space is said to be *totally null* if for every pair of vectors U^a and V^a tangent to Z at $x \in Z$, we have $U^a V_a = 0$. In particular, U^a and V^a must be null. It is not hard to convince oneself that the vectors U^a tangent to Z at x must *either* all have the form

$$U^a = \eta^A \pi^{A'}$$

where $\pi^{A'}$ depends only on Z and x *or* all have the form

$$U^a = \lambda^A \xi^{A'}$$

where λ^A depends only on Z and x. In the former case, Z is called an *α-surface*; in the latter case, a *β-surface*.

 Twistors for flat and curved space-time. *General relativity and gravitation, Vol. 2,* pp. 283–328, Plenum, New York and London, 1980

The importance of α surfaces is that their existence characterizes the vanishing of $\tilde{\Psi}_{A'B'C'D'}$. In fact, we have the following.

Theorem 4.1. Let M be a complex space–time. Then $\tilde{\Psi}_{A'B'C'D'} = 0$ is a necessary and sufficient condition that for every point $x \in M$ and every spinor $\pi_{A'}$ at x, there exists (locally) an α surface Z through x with tangent vectors of the form $\xi^{A}\pi^{A'}$ at x.

Proof. By Frobenius' theorem and spinor Ricci identities.

Suppose now that M is a right-conformally-flat space and define $\mathbb{P}\mathcal{T}$ as the space of α surfaces in M. Then, provided M is sufficiently "convex," $\mathbb{P}\mathcal{T}$ is a *three-dimensional complex manifold.* The reason for the convexity requirement is that if the space M has "holes" or "dents" in it, then $\mathbb{P}\mathcal{T}$ might, for example, fail to be Hausdorff. The restriction is intended merely to exclude such pathological possibilities. The three-dimensionality can be seen from the following argument: for every $x \in M$ (four dimensions) and spinor $\pi_{A'}$ up to proportionality (2 − 1 = one dimensional), there is a unique α-surface (which is two-dimensional), so the space of α-surfaces is 4 + 1 − 2 = three dimensional. One further remark: the structure of $\mathbb{P}\mathcal{T}$ is *conformally invariant*; in other words, g_{ab} and Ωg_{ab}, for any nowhere zero holomorphic function Ω, give rise to the same $\mathbb{P}\mathcal{T}$.

Conversely, $\mathbb{P}\mathcal{T}$ determines g_{ab} up to a conformal factor. To see this, it is convenient to use the theory of deformations of complex structure.[16,17] The argument runs as follows. Let $\mathcal{L}$ be some neighborhood of a line in flat projective twistor space $\mathbb{P}\mathcal{T}$. The complex projective lines in $\mathcal{L}$ correspond to the points in some region of complexified Minkowski space–time. Now deform the complex structure of $\mathcal{L}$ to yield a new complex manifold $\mathbb{P}\mathcal{T}$. This may not always be possible: for example, if $\mathcal{L}$ is the whole of $\mathbb{PT}$, then it cannot be deformed, since $\mathbb{CP}^3$ is *rigid.*[17] But we suppose here that $\mathcal{L}$ *is* deformable.

Theorem 4.2. As long as the deformation is sufficiently small, there exists a four-complex-parameter family of compact holomorphic curves (topology S^2) in $\mathbb{P}\mathcal{T}$, belonging to the same homology class as the lines in $\mathcal{L}$.

This theorem follows from a theorem of Kodaira.[18,19] We shall use it to construct a right-conformally-flat space M. The points of M are defined to be the holomorphic curves whose existence is guaranteed by Theorem 4.2. A conformal structure on M (i.e., a conformal class of metrics $\{\Omega g_{ab} | \Omega \text{ some function}\}$) is determined by giving a criterion for deciding whether or not two neighboring points of M are *null separated.* Our criterion is: two points in M are null separated iff the corresponding curves in $\mathbb{P}\mathcal{T}$ *intersect.*

Theorem 4.3.[16] This "conformal structure" is indeed quadratic, i.e., it arises from a complex metric g_{ab} on M.

Theorem 4.4. This conformal structure is right conformally flat.

Proof. Fix a point $\mathbf{Z}$ in $\mathbf{P}\mathcal{T}$ and consider all the holomorphic curves through $\mathbf{Z}$. These form a two-parameter family and hence correspond to a complex 2-surface in M. From the definition of null separation, we see immediately that this 2-surface is *totally null*. So it is either an α-surface or a β-surface. Here a question of convention arises, since we have not yet decided which are the primed spinors in M and which the unprimed spinors. We now fix the convention by saying that the above 2-surface is an α-surface. As $\mathbf{Z}$ varies over $\mathbf{P}\mathcal{T}$, we get plenty of α-surfaces in M, whence by Theorem 4.1 it follows that the conformal structure on M is right conformally flat. □

To sum up: Right-conformally-flat spaces (up to conformal rescaling) correspond to deformations of regions of $\mathbf{PT}$. Similarly, left-conformally-flat spaces correspond to deformations of regions of the dual twistor space $\mathbf{PT}^*$. To describe a *metric* g_{ab} (and not just a conformal metric) we need some additional structure on $\mathbf{P}\mathcal{T}$. It turns out that one can define additional structure in such a way that g_{ab} is necessarily *right flat*. The key to this structure is the following fact.

Theorem 4.5. Let M be a complex space–time. Then M is right flat $\Leftrightarrow$ there exists a covariantly constant primed spin frame $\Leftrightarrow$ the vector space of covariantly constant primed spinors is two-complex-dimensional.

To see what the implications of this theorem are for the twistor space structure, it is convenient to go to the *nonprojective* twistor space $\mathcal{T}$, which is defined as follows. Let M be right conformally flat and let Z be an α-surface in M. Then it is possible to find a primed spinor field $\pi_{A'}(x)$ on Z such that $\pi_{A'}$ is covariantly constant on Z and such that the tangent vectors at a point $x \in Z$ have the form $\lambda^A \pi^{A'}(x)$. The space of all such pairs $(Z, \pi_{A'})$ is the nonprojective twistor space $\mathcal{T}$ associated with M.

Now if M is right flat, then (in view of Theorem 4.5) the space of covariantly constant primed spinor fields on M is $\check{\mathbf{S}}^* = \mathbb{C}^2$. Since a constant spinor field on Z can be extended uniquely to a constant spinor field on M, we have a natural *fibering*

$$\mathcal{T} \to \mathcal{R} = \check{\mathbf{S}}^* - \{0\}$$

of the nonprojective twistor space. Conversely, if one is given such a fibering on $\mathcal{T}$, then it is possible[16] to define a *connection* on M compatible with its conformal structure. In other words we get a (complex) *Weyl structure* on M, which is necessarily right flat. For a *metric* on M we need more data on $\mathcal{T}$.

In what follows, all quantities are assumed appropriately homogeneous. Then the data that we require is (i) a nondegenerate 2-form τ on $\mathcal{B}$, and (ii) a nondegenerate 2-form μ on each fiber $\mathcal{F}_\pi$ of the fibering $\mathcal{T} \to \mathcal{B}$.

At this stage let us say what our additional structure corresponds to in flat twistor space. The fibering $\mathbb{T} \to \mathcal{B}$ is given by

$$(\omega^{A}, \pi_{A'}) \mapsto \pi_{A'} \tag{4.6}$$

and the forms τ and μ by

$$\begin{aligned} \tau &:= \tfrac{1}{2} I_{\alpha\beta}\, dZ^{\alpha} \wedge dZ^{\beta} = \tfrac{1}{2}\varepsilon^{A'B'}\, d\pi_{A'} \wedge d\pi_{B'} \\ \mu &:= \tfrac{1}{2}\varepsilon_{AB}\, d\omega^{A} \wedge d\omega^{B} \end{aligned} \tag{4.7}$$

Let us return now to the curved case. The holomorphic curves in $\mathbb{P}\mathcal{T}$ correspond to *holomorphic sections* of the bundle $\mathcal{T} \to \mathcal{B}$. We shall now show how to define a metric g_{ab} on M in terms of the twistor space structure.

Let $x \in M$ correspond to the holomorphic section $\mathbf{x}$ of $\mathcal{T}$; let y and z be neighboring points of x in M, and let $\mathbf{y}$ and $\mathbf{z}$ be the corresponding holomorphic sections of $\mathcal{T}$. The vectors $U = \overrightarrow{xy}$ and $V = \overrightarrow{xz}$ at $x \in M$ correspond to vectors $U(\pi)$ and $V(\pi)$ in $\mathcal{F}_\pi$, for each $\pi_{A'}$. For example, $U(\pi)$ is the vector in $\mathcal{F}_\pi$ from $\mathbf{x} \cap \mathcal{F}_\pi$ to $\mathbf{y} \cap \mathcal{F}_\pi$. Define

$$\mathcal{S}_\pi(U, V) := 2\mu\big(U(\pi), V(\pi)\big) \tag{4.8}$$

This is a function of $\pi_{A'}$, which is bilinear and skew symmetric in U and V. In flat space, by virtue of (4.7), we would have

$$\mathcal{S}_\pi(U, V) = U^{AA'} V_{A}{}^{B'} \pi_{A'} \pi_{B'}$$

Now suppose that U and V are *null* vectors. So by definition there exist spinors $\alpha_{A'}$ and $\beta_{A'}$ such that $U(\alpha) = V(\beta) = 0$. Then the inner product of U and V is defined to be

$$g(U, V) = \frac{\alpha^{A'} \beta_{A'}}{(\alpha^{B'} \pi_{B'})(\beta^{C'} \pi_{C'})} \mathcal{S}_\pi(U, V) \tag{4.9}$$

(this expression is in fact independent of $\pi_{A'}$). By linearity this defines a metric g_{ab} on M, which is compatible with the conformal structure of M. Furthermore,[16] we have the following.

Theorem 4.6. g_{ab} is right flat, and *every* right-flat metric (locally) can be constructed in this way.

4.2. Examples

We saw in the previous section that the general right-flat metric may be obtained by (i) deforming a region $\mathcal{R}$ of flat twistor space $\mathbb{T}$, preserving

certain structure; and then (ii) finding the holomorphic sections and hence the metric that this determines.

Part (i) of this program is easily carried out. For example, we could do it as follows. Cover $\mathscr{R}$ with two patches $\mathscr{Q}$ and $\hat{\mathscr{Q}}$. Assign coordinates $(\omega^A, \pi_{A'})$ to $\mathscr{Q}$ and $(\hat{\omega}^A, \hat{\pi}_{A'})$ to $\hat{\mathscr{Q}}$. $\mathscr{R}$ is determined by "patching" $\mathscr{Q}$ and $\hat{\mathscr{Q}}$ together according to

$$\hat{\omega}^A = \omega^A, \qquad \hat{\pi}_{A'} = \pi_{A'}$$

We obtain a *deformed* twistor space $\mathscr{T}$, with the required structure, by replacing this "flat-space" patching by

$$\begin{aligned} \hat{\omega}^A &= \omega^A + f^A(\omega^B, \pi_{B'}) \\ \hat{\pi}_{A'} &= \pi_{A'} \end{aligned} \tag{4.10}$$

where $f^A(Z^a)$ is holomorphic on $\mathscr{Q} \cap \hat{\mathscr{Q}}$, is homogeneous of degree 1, and satisfies

$$\det\left(\varepsilon_A{}^B + \frac{\partial f^B}{\partial \omega^A}\right) = 1 \tag{4.11}$$

(so as to preserve $\mu = \frac{1}{2}\varepsilon_{AB}\, d\omega^A \wedge d\omega^B$ on each fiber $\mathscr{F}_\pi$). Arranging for f^A to satisfy (4.11) is not difficult—it can, for example, be done by exponentiating a Hamiltonian vector field.

Part (ii) of the program—namely, that of finding the holomorphic sections of $\mathscr{T}$—is much more difficult to do explicitly. This problem is as yet unsolved, in general. But there are some special cases where it has been found possible to construct the holomorphic sections and hence the metric. These will now be described.

Let $p^{AA'\cdots C'} = p^{A(A'\cdots C')}$ be a constant spinor with one unprimed and $n > 0$ primed indices. Define $p^A := p^{AA'\cdots C'}\pi_{A'}\cdots\pi_{C'}$. Let $g(\gamma, \pi_{A'})$ be a holomorphic function of three complex variables with the homogeneity property

$$g(\lambda^{n+1}\gamma, \lambda\pi_{A'}) = \lambda^{1-n} g(\gamma, \pi_{A'}) \tag{4.12}$$

for all nonzero $\lambda \in \mathbb{C}$. Then take the patching function f^A to be

$$f^A(\omega^B, \pi_{B'}) = p^A g(p_B\omega^B, \pi_{B'}) \tag{4.13}$$

It is easy to check that f^A is homogeneous of degree 1 and satisfies (4.11). The implication of this particular choice of patching function is that the function $p_B\omega^B$ is a *globally defined* holomorphic function on the deformed twistor space $\mathscr{T}$, homogeneous of degree $n + 1$. This is because, by (4.10) and (4.13), $\hat{p}_B\hat{\omega}^B = p_B\omega^B$. In general $\mathscr{T}$ would *not* admit such functions: the only global homogeneous functions on $\mathscr{T}$ would be those lifted from the base $\mathscr{B}$. So the patching functions (4.13) give some, but not all, right-flat spaces.

Now that we have $\mathcal{T}$, determined by (4.10) and (4.13), can we find the holomorphic sections? One way of doing so is as follows. First we "split" the function $g(ip_{A}x^{AA'}\pi_{A'}, \pi_{A'})$:

$$g(ip_{A}x^{AA'}\pi_{A'}, \pi_{A'}) = h(x, \pi_{A'}) - \hat{h}(x, \pi_{A'})$$

where $h(x, \pi_{A'})$ is holomorphic for $(ix^{AA'}\pi_{A'}, \pi_{A'}) \in \mathcal{Q}$ and $\hat{h}$ for $(ix^{AA'}\pi_{A'}, \pi_{A'}) \in \hat{\mathcal{Q}}$. Now $g(ip_{A}x^{AA'}\pi_{A'}, \pi_{A'})$ is, by (4.12), homogeneous of degree $1 - n$ in $\pi_{A'}$ and so in order for such a splitting to be possible we must have [cf. (2.83)]

$$1 - n \geq -1$$

$$\Leftrightarrow n \leq 2$$

Supposing this to be the case, we can write down the holomorphic sections of $\mathcal{T}$ immediately: they are given by

$$\omega^{A}(x, \pi_{B'}) = ix^{AA'}\pi_{A'} + p^{A}h(x, \pi_{B'}) \qquad \text{in } \mathcal{Q}$$

$$\hat{\omega}^{A}(x, \pi_{B'}) = ix^{AA'}\pi_{A'} + p^{A}\hat{h}(x, \pi_{B'}) \qquad \text{in } \hat{\mathcal{Q}}$$

If $n > 2$, this procedure breaks down and it is not yet known how to construct the holomorphic sections.

For $n = 0, 1, 2$ one can now go ahead and compute the corresponding right-flat metrics. The details of the computation have appeared elsewhere[20] and we shall merely state the results here.

Case n = 0. Here the metric can be written as

$$ds^2 = du\,dv + dx\,dy + f(v, x)\,dv^2$$

where $f(v, x)$ is an arbitrary function. This metric is of algebraic type $[4] \otimes [-]$ and may be thought of as an anti-self-dual *pp* wave.[21]

Case n = 1. First the function g is used to define a field $\psi_{A}(x)$ by

$$\psi_{A}(x) := \oint (2\pi i \pi_{0'})^{-1} \dot{g}(ip_{B}x^{BB'}\pi_{B'}, \pi_{B'}) p_{A}{}^{A'}\pi_{A'}\pi_{D'}\,d\pi^{D'}$$

where $\dot{g}(\gamma, \pi_{A'}) := \partial g(\gamma, \pi_{A'})/\partial\gamma$. Notice that ψ_{A} is a static neutrino field, i.e.,

$$\nabla^{AA'}\psi_{A} = 0, \qquad p^{AA'}\nabla_{AA'}\psi_{B} = 0$$

We may suppose without loss of generality that $p_{a}p^{a} = 1$. Put

$$t := p_{a}x^{a}$$

$$V := 1 + p_{0'}^{A}\psi_{A}$$

$$\xi := (V - 1)\,dt + \psi_{A}\,dx^{A0'}$$

Then the metric turns out to be

$$ds^2 = V^{-1}(dt + \xi)^2 - V\delta_{jk}\,dx^{j}\,dx^{k}, \qquad j, k = 1, 2, 3$$

These metrics clearly possess the Killing vector $\partial/\partial t$. They include, for example, the "self-dual Taub–NUT" solution.[22]

Case $n = 2$. Take $p^{AA'B'} = o^A o^{A'} o^{B'} + \iota^A \iota^{A'} \iota^{B'}$ and define

$$\phi_{A'B'} := (2\pi i)^{-1} \oint \pi_{A'} \pi_{B'} \dot{g}(ip_C x^{CC'} \pi_{C'}, \pi_{C'}) \pi_{D'} \, d\pi^{D'}$$

$$\theta := 1 - \phi_{0'1'}$$

$$\phi := \phi_{0'0'}$$

$$\tilde{\phi} := \phi_{1'1'}$$

$$\Lambda := \theta^2 - \phi\tilde{\phi}$$

Then in the coordinate system $y^a = (x^{00'}, x^{11'}, x^{01'}, x^{10'})$, the metric is

$$g^{ab} = -\Lambda^{-1} \begin{bmatrix} 2\phi & -\theta & 0 & \tilde{\phi} \\ -\theta & 2\tilde{\phi} & \phi & 0 \\ 0 & \phi & 0 & \theta \\ \tilde{\phi} & 0 & \theta & 0 \end{bmatrix}$$

These metrics do not in general appear to have any Killing vectors, and they are algebraically general. Nevertheless, since they depend on an arbitrary function of two variables (namely, g modulo its homogeneity), they are still a long way from being the *general* right-flat solution, which would require an arbitrary function of three variables.

4.3. Massless Fields on a Half-Flat Background

The previous section was concerned with using "curved twistor space" theory to find solutions of Einstein's equations. Another problem of interest is whether aspects of flat-space twistor theory (such as the procedure for solving the massless free-field equations) can be extended to curved space-times. We shall see that such an extension *does* exist for right-flat spaces. Carrying out our contour integrals on the deformed twistor space will give rise to (test) fields which interact in the "correct" way with the right-flat background space.

First let us remark on the algebraic restrictions known as Buchdahl constraints, which arise in connection with solutions of the massless field equation

$$\nabla^{AA'} \phi_{\underbrace{AB\cdots D}_{2s}} = 0 \tag{4.14}$$

in curved space-time. By differentiating (4.14), commuting the derivatives and using the spinor Ricci identities, we can derive, for $s \geq \frac{3}{2}$,

$$\Psi^{ABC}{}_{(D} \phi_{E\cdots G)ABC} = 0 \tag{4.15}$$

Similarly,

$$\nabla^{AA'}\phi_{\underbrace{A'B'\cdots D'}_{2s}} = 0 \tag{4.16}$$

implies

$$\tilde{\Psi}^{A'B'C'}{}_{(D'}\phi_{E'\cdots G')A'B'C'} = 0 \tag{4.17}$$

for $s \geq \frac{3}{2}$. In a right-flat space, $\tilde{\Psi}_{A'B'C'D'}$ vanishes and the condition (4.17) is void, but (4.15) is nontrivial. So we might expect to be able to solve (4.16), the wave equation $\Box\phi = 0$, and (4.14) with $s \leq 1$ on a right-flat background. On the other hand, in order to have as many solutions as in flat space, the equation (4.14) with $s > 1$ would need to be modified in such a way that no algebraic constraints arise. As we shall see, and as was first noticed by G. A. J. Sparling, twistor theory suggests such a modification and provides the solutions of the modified equations.

Let M be a right-flat space–time and $\mathcal{T}$ the corresponding deformed twistor space. Let f be a holomorphic function on $\mathcal{T}$, homogeneous of degree $-2s - 2$, with $s \geq 0$. Assuming that f has a suitable singularity structure, we want to perform a contour integral of f to extract a massless field. The first step (cf. Section 2.2) is to "put the twistor Z through the space–time point x." In our case, this can be achieved as follows: a pair $(x, \pi_{A'})$, where $x \in M$ and $\pi_{A'}$ is a constant spinor, determines a nonprojective twistor which we denote $Z_{x,\pi}$. Now define a function $F(x, \pi)$ by

$$F(x, \pi) := f(Z_{x,\pi})$$

It is clear from the definition that

$$\begin{aligned} &\lambda^A \pi^{A'} \nabla_{AA'} F(x, \pi) = 0 \qquad \text{for all } \lambda^A \\ &\Rightarrow \pi^{A'} \nabla_{AA'} F(x, \pi) = 0 \end{aligned} \tag{4.18}$$

Put

$$\phi_{A'\cdots C'}(x) = \frac{1}{2\pi i} \oint \underbrace{\pi_{A'} \cdots \pi_{C'}}_{2s} F(x, \pi) \pi_{D'}\, d\pi^{D'}$$

It follows immediately from (4.18) that $\phi_{A'\cdots C'}$ is a solution of (4.16), or of $\Box\phi = 0$ if $s = 0$. This procedure can easily be "cohomologized": an element of $H^1(\mathbb{P}\mathcal{T}, \mathcal{O}(-2s - 2))$ will give rise to massless spin-s field in the usual way.

Now take f to be homogeneous of degree $2s - 2$, with $s > 0$. To construct a space–time field we use the splitting procedure. For each fixed x, split $F(x, \pi)$:

$$F(x, \pi) = \hat{h}(x, \pi) - h(x, \pi) \tag{4.19}$$

This can always be done [cf. (2.83)], since $2s - 2 \geq -1$ whenever $s > 0$. Operating on (4.19) with $\pi^{A'}\nabla_{AA'}$ and using (4.18) gives

$$\pi^{A'}\nabla_{AA'}\hat{h} = \pi^{A'}\nabla_{AA'}h$$

from which it follows that $\pi^{A'}\nabla_{AA'}h$ must be a *homogeneous polynomial* of degree $2s-1$ in $\pi_{A'}$:

$$\pi^{A'}\nabla_{AA'}h(x,\pi) = \Phi_{\underbrace{AA'\cdots C'}_{2s-1}}(x)\pi^{A'}\cdots\pi^{C'} \tag{4.20}$$

where $\Phi_{AA'\cdots C'} = \Phi_{A(A'\cdots C')}$ is some spinor field on M, defined uniquely by h. Operating on (4.20) with $\pi^{D'}\nabla_{D'}{}^{A}$ leads to

$$\nabla_{A(A'}\Phi^{A}_{B'\cdots D')} = 0 \tag{4.21}$$

For $s = \frac{1}{2}$, equation (4.21) reduces to

$$\nabla^{AA'}\Phi_{A} = 0$$

For $s \geq 1$, $\Phi^{A}_{A'\cdots C'}$ should be thought of as a *left-handed potential* for a massless field. The freedom in the splitting (4.19) is

$$h(x,\pi) \mapsto h(x,\pi) + \lambda_{A'\cdots C'}(x)\pi^{A'}\cdots\pi^{C'}$$

$$\hat{h}(x,\pi) \mapsto \hat{h}(x,\pi) + \lambda_{A'\cdots C'}(x)\pi^{A'}\cdots\pi^{C'}$$

substituting this into (4.20) gives a *gauge freedom* in $\Phi_{AA'\cdots D'}$, namely,

$$\Phi_{AA'B'\cdots D'} \mapsto \Phi_{AA'B'\cdots D'} + \nabla_{A(A'}\lambda_{B'\cdots D')} \tag{4.22}$$

A spin-s field $\phi_{A\cdots D}$ can be defined in terms of the potential $\Phi_{AA'\cdots C'}$ by

$$\phi_{A\cdots D} := \underbrace{\nabla^{A'}_{(A}\cdots\nabla^{C'}_{C}}_{2s-1 \text{ derivatives}}\Phi_{D)A'\cdots C'} \tag{4.23}$$

In the case $s = 1$, ϕ_{AB} defined by (4.23) is gauge invariant and automatically satisfies the Maxwell equations $\nabla^{AA'}\phi_{AB} = 0$; this follows from (4.21), after some algebra. When $s \geq \frac{3}{2}$, however, the Buchdahl constraints come into play and this nice state of affairs no longer holds. If $\Psi_{ABCD} \neq 0$, then $\phi_{A\cdots D}$ is not gauge invariant and the equation it satisfies is a modified version of (4.14). As an illustration, consider the case $s = \frac{3}{2}$. Using the spinor Ricci identities and equation (4.21), one can show that (i) under the gauge transformation (4.22), the field ϕ_{ABC} transforms as

$$\phi_{ABC} \mapsto \phi_{ABC} - \tfrac{1}{2}\Psi_{ABCD}\nabla^{DA'}\lambda_{A'} \tag{4.24}$$

and (ii) the field ϕ_{ABC} satisfies

$$\nabla^{AA'}\phi_{ABC} = -\Psi_{ABCD}\nabla^{B'A}\Phi^{DA'}_{B'} \tag{4.25}$$

It is easily checked that the field equations (4.25) are invariant under the gauge transformations (4.22) and (4.24); and the very fact that we can generate solutions of them using twistor functions indicates that they are not subject to algebraic constraints of the Buchdahl type.

 Twistors for flat and curved space-time. *General relativity and gravitation, Vol. 2,* pp. 283–328, Plenum, New York and London, 1980

5. Hypersurface Twistors

5.1. CR Structure and Hypersurface Twistors

It was suggested in Section 3 that, in a general curved space-time M, twistors can be defined relative to some hypersurface $\mathscr{S}$. This can indeed be done and the resulting *hypersurface twistor* formalism is at present the best way we know of generalizing twistor theory to space-times which are not half conformally flat. It applies only when both M and $\mathscr{S}$ are real analytic, but the role played by this analyticity condition is not very crucial; it seems likely that the theory could be modified so as to avoid it.

The basic idea is as follows. Suppose for the moment that $\mathscr{S}$ is a (spacelike) Cauchy surface. Then (as we shall see presently) the space $\mathbb{P}\mathcal{N}$ of unscaled null geodesics in M acquires, with respect to $\mathscr{S}$, a *CR structure*.[15] Let $\mathbb{P}\mathcal{N}(\mathscr{S})$ denote $\mathbb{P}\mathcal{N}$ equipped with this structure. The CR structure is linked to the geometry of M via the following:

Hypersurface Kerr Theorem. Let $\pi_{A'}$ be an analytic spinor field on $\mathscr{S}$. Then $\pi_{A'}$ is shear free on $\mathscr{S}$ iff it is given locally by an equation $f = 0$, where f is a solution of the tangential Cauchy–Riemann equations on $\mathbb{P}\mathcal{N}(\mathscr{S})$. It follows from this theorem and from the fact that conformal curvature generates shear, that if $\mathscr{S}$ and $\mathscr{S}'$ are two distinct hypersurfaces in M, then $\mathbb{P}\mathcal{N}(\mathscr{S})$ and $\mathbb{P}\mathcal{N}(\mathscr{S}')$ are in general *different*. The CR structure *shifts* as one moves from $\mathscr{S}$ to $\mathscr{S}'$

The CR structure of $\mathbb{P}\mathcal{N}(\mathscr{S})$ can be defined in either of two equivalent ways: one could write down the tangential Cauchy–Riemann equations, or one could construct a three-dimensional complex manifold containing $\mathbb{P}\mathcal{N}(\mathscr{S})$ as a real hypersurface. We shall describe the second method, by constructing a hypersurface twistor space $\mathbb{P}\mathcal{T}(\mathscr{S})$.

The first step is to complexify $\mathscr{S}$ and M, so that we get a complex hypersurface $\mathbb{C}\mathscr{S}$ contained in a complex space-time $\mathbb{C}M$. If M were Minkowski space-time, then a twistor Z^α could be represented by the pair $(Z, \pi_{A'})$ where Z is an α-plane. The intersection $Z \cap \mathbb{C}\mathscr{S}$ is a complex curve ζ in $\mathbb{C}\mathscr{S}$, with tangent vector proportional to

$$v^a = t^{AB'}\pi_{B'}\pi_{A'} \tag{5.1}$$

Here t^a is a vector normal to $\mathscr{S}$. One may regard the spinor $\pi_{A'}$ as being parallelly propagated along ζ, i.e.,

$$v^a\nabla_a\pi_{C'} = t^{AB'}\pi_{B'}\pi^{A'}\nabla_{AA'}\pi_{C'} = 0 \tag{5.2}$$

This is all very well for flat space-time, but in general curved space-times we do not have α-surfaces. We do, however, still have pairs $(\zeta, \pi_{A'})$, where (a) ζ

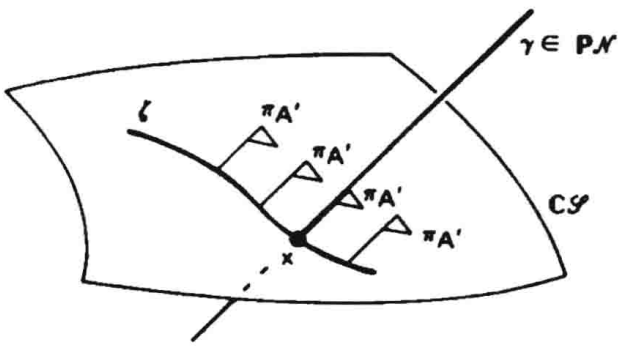


Fig. 7. x is a real point, so that (ζ, π_A) is null.

is a complex curve in $\mathbb{C}\mathcal{S}$, (b) $\pi_{A'}$ is a spinor field on ζ satisfying (5.2), (c) the tangent vector to ζ has the form (5.1). Such pairs are, by definition, *hypersurface twistors*. The space $\mathcal{T}(\mathcal{S})$ of hypersurface twistors is a four-dimensional-complex manifold.

Factoring out by the equivalence relation

$$(\zeta, \pi_{A'}) \sim (\zeta, \lambda\pi_{A'}), \qquad \lambda \in \mathbb{C}, \lambda \neq 0$$

takes us from $\mathcal{T}(\mathcal{S})$ to the *projective* hypersurface twistor space $\mathbb{P}\mathcal{T}(\mathcal{S})$. A twistor $(\zeta, \pi_{A'})$ is said to be *null* if the curve ζ contains a real point (in flat space–time this is equivalent to the usual definition). The space of projective null twistors is precisely $\mathbb{P}\mathcal{N}(\mathcal{S})$ (see Figure 7) and this is what we set out to achieve. It may be remarked that the structure of $\mathbb{P}\mathcal{T}(\mathcal{S})$, $\mathcal{T}(\mathcal{S})$, and $\mathbb{P}\mathcal{N}(\mathcal{S})$ is *conformally invariant.*[9]

The greatest drawback of the hypersurface twistor description is its dependence on the choice of hypersurface $\mathcal{S}$. It seems that the formalism will be most useful when there is a natural choice of hypersurface. For example, if M is asymptotically flat, we can take $\mathcal{S} = \mathcal{I}^+$ or $\mathcal{I}^-$ (see Section 5.2); if M is static, we could take $\mathcal{S}$ to be orthogonal to the timelike Killing vector. These possibilities have yet to be fully investigated and exploited.

5.2. Twistors Relative to a Null Hypersurface

In this section we examine the hypersurface twistor space $\mathcal{T}(\mathcal{S})$ in the case where $\mathcal{S}$ is *null*. The fact that $\mathcal{S}$ is no longer transverse to all null geodesics (as was true when $\mathcal{S}$ was spacelike) does introduce some complications. However, these are not serious and they are compensated for by some extra structure (in particular, a *Kähler* structure) which the twistor space $\mathcal{T}(\mathcal{S})$ acquires.

If $\mathcal{S}$ is null, then its normal vector t^a has the form

$$t^a = \xi^A\bar{\xi}^{A'}$$

Substituting this into (5.1) gives $v^a = (\bar{\xi}^{B'}\pi_{B'})\xi^A\pi^{A'}$, which vanishes when $\pi^{A'}$ is proportional to $\bar{\xi}^{A'}$. So it is preferable to omit the factor $\bar{\xi}^{B'}\pi_{B'}$ and to define a hypersurface twistor (with respect to a *null* hypersurface $\mathscr{S}$) as a pair $(\zeta, \pi_{A'})$, where (a) ζ is a holomorphic curve in $\mathbb{C}\mathscr{S}$ with tangent vector of the form $\xi^A\pi^{A'}$, and (b) $\pi_{A'}$ is a spinor field on ζ satisfying $\xi^A\pi^{A'}\nabla_{AA'}\pi_{B'} = 0$.

Projective hypersurface twistors and null hypersurface twistors are defined exactly as in the spacelike-$\mathscr{S}$ case. Similarly, one can define *conjugate* hypersurface twistor space $\check{\mathscr{T}}(\mathscr{S})$ by simply interchanging primed and unprimed indices in the above definition.

The primary application of this formalism is that of taking $\mathscr{S}$ to be the past or future null infinity ($\mathscr{I}^-$ or $\mathscr{I}^+$) of an *asymptotically flat* space–time M.[6.13] Asymptotic flatness means essentially that it is possible, after making a conformal rescaling of the space–time metric in M, to attach, in a suitably smooth way, the null hypersurfaces $\mathscr{I}^-$ and $\mathscr{I}^+$ as boundaries to M. Since the hypersurface twistor space construction is conformally invariant, it is unaffected by conformal rescaling. Whereas the conformal curvature spinor Ψ_{ABCD} *vanishes* on $\mathscr{I}$, its first covariant derivative at $\mathscr{I}$ (with respect to the connection of the rescaled metric) does not, in general. This has the implication that the twistor spaces $\mathbb{P}\mathscr{T}(\mathscr{I}^+)$ and $\mathbb{P}\mathscr{T}(\mathscr{I}^-)$, called the future and past (projective) *asymptotic* twistor spaces of M, respectively, are generally *curved*, in the sense that they cannot be holomorphically imbedded in $\mathbb{CP}^3$. In fact, it is these derivatives of Ψ_{ABCD} on $\mathscr{I}^+$ and on $\mathscr{I}^-$ that measure the outgoing and incoming gravitational radiation fields, respectively, for the space–time M. This information is coded into the structure of $\mathbb{P}\mathscr{T}(\mathscr{I}^+)$ and $\mathbb{P}\mathscr{T}(\mathscr{I}^-)$, or, equivalently, into the structure of the nonprojective asymptotic twistor spaces $\mathscr{T}(\mathscr{I}^\pm)$. In particular, if the gravitational radiation in M is purely retarded, then $\mathscr{T}(\mathscr{I}^-)$ is flat.

Returning to the case of a general null hypersurface $\mathscr{S}$, we find that the twistor space possesses some structure that if did not have in the spacelike case. This is a "*scalar product*," meaning a holomorphic function Π on an open subset $\mathscr{Q}$ of $\mathscr{T}(\mathscr{S}) \times \check{\mathscr{T}}(\mathscr{S})$, homogeneous of degree 1 on each of $\mathscr{T}(\mathscr{S})$ and $\check{\mathscr{T}}(\mathscr{S})$:

$$\Pi : \mathscr{Q} \to \mathbb{C}$$

with

$$\Pi(\lambda Z, \mu\check{Z}) = \lambda\mu\,\Pi(Z, \check{Z})$$

where

$$Z \in \mathscr{T}(\mathscr{S}), \qquad \check{Z} \in \check{\mathscr{T}}(\mathscr{S}), \qquad \lambda, \mu \in \mathbb{C} - \{0\}$$

For the special case $\mathscr{S} = \mathscr{I}^\pm$, the definition of scalar product is given in References 7 and 9. The definition for a general null $\mathscr{S}$ is essentially the same. Let ζ and $\check{\zeta}$ be the twistor curve and dual twistor curve on $\mathbb{C}\mathscr{S}$ for Z

and $\tilde{Z}$, respectively. Define the region $\mathbb{P}\mathcal{Q}$ to consist of those pairs $(\zeta, \tilde{\zeta})$ for which a complex null geodesic generator η of $\mathcal{S}$ meets both ζ and $\tilde{\zeta}$ (and suppose, for simplicity, that this generator is always simply connected). On dimensional and continuity grounds we would expect $\mathbb{P}\mathcal{Q}$ to be an open subset of $\mathbb{P}\mathcal{T}(\mathcal{S}) \times \mathbb{P}\tilde{\mathcal{T}}(\mathcal{S})$. The set $\mathcal{Q}$ is the corresponding subset of $\mathcal{T}(\mathcal{S}) \times \tilde{\mathcal{T}}(\mathcal{S})$. To define $\Pi(Z, \tilde{Z})$ we consider $(0, \pi_{A'})$ as a local twistor at the point $\zeta \cap \eta$ (representing Z) and $(\tilde{\pi}_A, 0)$ as a dual local twistor at $\tilde{\zeta} \cap \eta$, then carry one of these local twistors along some path in η which connects $\zeta \cap \eta$ to $\tilde{\zeta} \cap \eta$, using local twistor transport, and finally take its local twistor scalar product with the other.

A *local twistor*[7] and a *dual local twistor* at a point p of $\mathbb{C}M$ are each described, respectively, by a pair of spinors at p:

$$(\omega^A, \pi_{A'}), (\tilde{\pi}_A, \tilde{\omega}^{A'}) \tag{5.3}$$

This holds if the metric g_{ab} of $\mathbb{C}M$ is fixed, but under the conformal rescaling

$$g_{ab} \mapsto \Omega^2 g_{ab}; \qquad \varepsilon_{AB} \mapsto \Omega\varepsilon_{AB}, \qquad \varepsilon_{A'B'} \mapsto \Omega\varepsilon_{A'B'} \tag{5.4}$$

we must make the corresponding replacements

$$(\omega^A, \pi_{A'}) \mapsto (\omega^A, \pi_{A'} + i\Upsilon_{AA'}\omega^A), (\tilde{\pi}_A, \tilde{\omega}^{A'}) \mapsto (\tilde{\pi}_A - i\Upsilon_{AA'}\tilde{\omega}^{A'}, \tilde{\omega}^{A'}) \tag{5.5}$$

where

$$\Upsilon_a = \Omega^{-1}\nabla_a\Omega = \nabla_a \log \Omega$$

Local twistor transport carries $(\omega^A, \pi_{A'})$ and $(\tilde{\pi}_A, \tilde{\omega}^{A'})$ along some path in $\mathbb{C}M$, having tangent vector u^a, according to

$$\begin{aligned} u^b\nabla_b\omega^A &= -iu^{AB'}\pi_{B'}, \qquad u^b\nabla_b\pi_{A'} = -iP_{ABA'B'}u^{BB'}\omega^A, \\ u^b\nabla_b\tilde{\pi}_A &= iP_{ABA'B'}u^{BB'}\tilde{\omega}^{A'}, \qquad u^b\nabla_b\tilde{\omega}^{A'} = iu^{BA'}\tilde{\pi}_B \end{aligned} \tag{5.6}$$

where

$$P_{ABA'B'} = \Phi_{ABA'B'} - \Lambda\varepsilon_{AB}\varepsilon_{A'B'} = -\tfrac{1}{2}R_{ab} + \tfrac{1}{12}Rg_{ab}$$

Under the transformation law (5.4), (5.5) is conformally invariant. Note also that the forms $(0, \pi_{A'})$ and $(\tilde{\pi}_A, 0)$ are preserved, respectively, along ζ and along $\tilde{\zeta}$. The local twistor scalar product can be evaluated at any *one* point p and is

$$(\omega^A, \pi_{A'}) \cdot (\tilde{\pi}_A, \tilde{\omega}^{A'}) = \omega^A\tilde{\pi}_A + \pi_{A'}\tilde{\omega}^{A'}$$

this being conformally invariant and preserved under local twistor transport. At any *real* point $p \in M$, the complex conjugate of $(\omega^A, \pi_{A'})$ is the dual local twistor $(\bar{\pi}_A, \bar{\omega}^{A'})$.

In flat space-time, $\Pi(Z, \tilde{Z})$ reduces to the usual bilinear scalar product $Z^a\bar{Z}_a$. But in the general case Π has only very limited linearity properties. If Z and $\tilde{Z}$ are varied in such a way that the line Π is held fixed, then Π *can* be said to maintain its bilinear form—referring everything to the linear local

twistor structure along η. But if η varies, this local twistor structure varies in a nonintegrable way, owing to the nonintegrability of (5.5) when Ψ_{ABCD} or its first covariant derivative is nonzero on $\mathscr{S}$.

Combining Π with the conjugation map

$$\mathscr{T}(\mathscr{S}) \ni Z \mapsto \bar{Z} \in \bar{\mathscr{T}}(\mathscr{S})$$

gives a "norm" $\Pi(Z, \bar{Z})$ on some suitable neighborhood of $\mathscr{N}(\mathscr{S})$ in $\mathscr{T}(\mathscr{S})$, with the property that Z is null iff $\Pi(Z, \bar{Z}) = 0$. Thinking of $\Pi(Z, \bar{Z})$ as a "Kähler scalar" leads to a *Kähler structure* on $\mathscr{T}(\mathscr{S})$,[17] i.e., a second derivative of Π, once holomorphic and once antiholomorphic, gives the Kähler metric. The Kähler curvature can be expressed in terms of the space–time curvature. [See Reference 23 for the case $\mathscr{S} = \mathscr{I}^{\pm}$; outgoing gravitational radiation then gives rise to Kähler-curvature for $\mathscr{T}(\mathscr{I}^{+})$ and incoming radiation to that for $\mathscr{T}(\mathscr{I}^{-})$.]

As was pointed out in Section 4.1, a curved twistor space $\mathbb{P}\mathscr{T}(\mathscr{S})$ has associated with it a four-complex-dimensional "space–time," with a conformal metric which is right conformally flat. Let us denote this complex space–time by $\mathscr{H}(\mathscr{S})$. The points of $\mathscr{H}(\mathscr{S})$ are holomorphic curves in $\mathbb{P}\mathscr{T}(\mathscr{S})$ and each point of $\mathbb{P}\mathscr{T}(\mathscr{S})$ corresponds to a curve ζ in $\mathbb{C}\mathscr{S}$; so points of $\mathscr{H}(\mathscr{S})$ can also be thought of as complex 2-surfaces in $\mathbb{C}\mathscr{S}$ which are ruled by hypersurface twistor curves. In the case $\mathscr{S} = \mathscr{I}^{+}$, the ζ curves and $\tilde{\zeta}$ curves turn out to be the complex *null geodesics* on $\mathbb{C}\mathscr{I}^{+}$. The 2-surfaces ruled by $\tilde{\zeta}$ curves are then often called "good cuts" or "shear-free cuts" of $\mathbb{C}\mathscr{I}$. They were first introduced by Newman and his colleagues from a somewhat different point of view. In fact Newman's $\mathscr{H}$ space is what we would label here as $\check{\mathscr{H}}(\mathscr{I}^{+})$ [holomorphic curves in $\bar{\mathscr{T}}(\mathscr{I}^{+})$]. It was Newman's $\mathscr{H}$-space construction that largely motivated the original definition of asymptotic twistor space, which, in turn led to the more general concept of hypersurface twistor space. For a recent review of the theory of $\mathscr{H}$ space, see Reference 3.

It should be pointed out, however, that there is a certain qualitative distinction between those 2-surfaces which give rise to points of $\mathscr{H}(\mathscr{S})$ when $\mathscr{S}$ is spacelike and those when $\mathscr{S}$ is null. Consider the spacelike case first. Suppose that p is any point of $\mathscr{S}$ and consider the (holomorphic) system P of ζ curves passing through p. There is exactly one such ζ curve orthogonal to every real null direction at p, so the topology of P is compact, namely, S^2. The same topology for P is maintained if p is any point of $\mathbb{C}\mathscr{S}$ not too far from $\mathscr{S}$. Thus we have a three-complex-parameter family of such loci P (compact holomorphic curves) in $\mathbb{P}\mathscr{T}(\mathscr{S})$. This family constitutes a complex 3-surface in $\mathscr{H}(\mathscr{S})$ in natural 1 : 1 correspondence with the points of $\mathbb{C}\mathscr{S}$ (or, at least, those points of $\mathbb{C}\mathscr{S}$ not too far from $\mathscr{S}$). Now, moving to a general point of $\mathscr{H}(\mathscr{S})$ close to this 3-surface, we find that it is represented by a locus P' in $\mathbb{P}\mathscr{T}(\mathscr{S})$ that does not differ greatly from those of the type P, just considered. The 2-surface, ruled by ζ curves, that corresponds to P' becomes, in the limit when P' approaches P, the cone of ζ curves with

vertex p. In this sense, the 2-surfaces in question remain somewhat "local" in $\mathbb{C}\mathcal{S}$.

When $\mathcal{S}$ is null, however, the construction degenerates for $p \in \mathcal{S}$. There is a unique ζ curve through p, as before, orthogonal to each real null direction at p *other than* that of the normal $\xi^A\bar{\xi}^{A'}$ at p. This normal is orthogonal to all ζ curves through p and we do not have a clear way to complete the S^2 family of ζ curves through p in a nonsingular fashion. The generator of $\mathcal{S}$ through p may itself be considered to be a ζ curve. However, even in the case when M is Minkowski space and $\mathcal{S} = \mathcal{I}^+$ this does not quite provide us with the correct region **I** of **PT** but gives the "blown-up" projective twistor space instead.[7] The question of how best to define the hypersurface twistors corresponding to the normal directions in a null $\mathcal{S}$ is a delicate one and we shall not discuss it here. In any case the situation does not closely resemble the one considered above for a spacelike $\mathcal{S}$. We cannot adequately obtain the appropriate systems of ζ lines (or 2-surfaces generated by them) representing points of $\mathcal{H}(\mathcal{S})$ as being "close" to the system of ζ lines through a point of $\mathcal{S}$. Instead, it is more appropriate to think of the required 2-surfaces of ζ lines as being "close" to $\mathbb{C}\mathcal{S} \cap \mathbb{C}\mathcal{K}$, where $\mathbb{C}\mathcal{K}$ is the complexified light cone of a point of M *not* on $\mathcal{S}$. (In the case when M is Minkowski space this gives such 2-surfaces exactly.) Alternatively, we may view the 2-surfaces as, effectively, cross sections of $\mathcal{S}$, the null generators of $\mathcal{S}$ providing $\mathcal{S}$ with a fiber space structure.

In the particular situations $\mathcal{S} = \mathcal{I}^\pm$ these generators of $\mathcal{S}$ are shear-free.[6] This turns out to have the consequence that the induced conformal metric of each cross section, or "cut," of $\mathbb{C}\mathcal{I}^\pm$ (whether generated by ζ lines or not) is the "same" in the sense that it mapped without change along the generators of $\mathbb{C}\mathcal{I}^\pm$. Hence the two systems of isotropic lines on any cross section map along these generators to isotropic lines on the other cross sections. In other words, these isotropic lines lie on two systems of (totally null) planes on $\mathcal{I}^\pm$, which we refer to as the α-planes and the β-planes on $\mathcal{I}^\pm$. Every ζ curve lies on a β-plane on $\mathbb{C}\mathcal{I}^\pm$ (defined, in the case of Minkowskian M, by $W_\alpha = I_{\alpha\beta}Z^\beta$) and every $\tilde{\zeta}$ curve on an α-plane on $\mathbb{C}\mathcal{I}^\pm$ (see Figure 8). Thus there is a canonical projection from the space of ζ curves to the space of β-planes on $\mathcal{I}^\pm$, i.e., from $\mathbb{P}\mathcal{T}(\mathcal{I}^\pm)$ to a $\mathbb{CP}^1$. This gives us a projection also from $\mathcal{T}(\mathcal{I}^\pm)$ to $\mathcal{B} = \mathbb{C}^2 - \{0\}$. The "strong conformal geometry"[6] of $\mathcal{I}^\pm$ enables the remaining additional structure described in Section 4.2 to be assigned to $\mathcal{T}$ and we have (a) a fibering $\mathcal{T} \to \mathcal{B}$; (b) a 2-form τ on $\mathcal{B}$; and (c) a 2-form μ on each fiber (appropriately homogeneous). Thus, as in Section 4.1, the $\mathcal{H}$ space $\mathcal{H}(\mathcal{I}^\pm)$ possesses a canonically defined right-flat metric.

In fact,[3] the curvature Ψ_{ABCD} of $\mathcal{H}(\mathcal{I}^-)$ can be regarded as arising precisely from the anti-self-dual part of the incoming gravitational radiation for M and that on $\tilde{\mathcal{H}}(\mathcal{I}^-)$ from the self-dual part. Likewise, $\mathcal{H}(\mathcal{I}^+)$ and $\tilde{\mathcal{H}}(\mathcal{I}^+)$ have curvatures arising from the anti-self-dual and self-dual parts, respectively, of the outgoing gravitational radiation. Thus twistor methods provide a natural means of describing these two parts of the gravitational

 Twistors for flat and curved space-time. *General relativity and gravitation, Vol. 2,* pp. 283–328, Plenum, New York and London, 1980

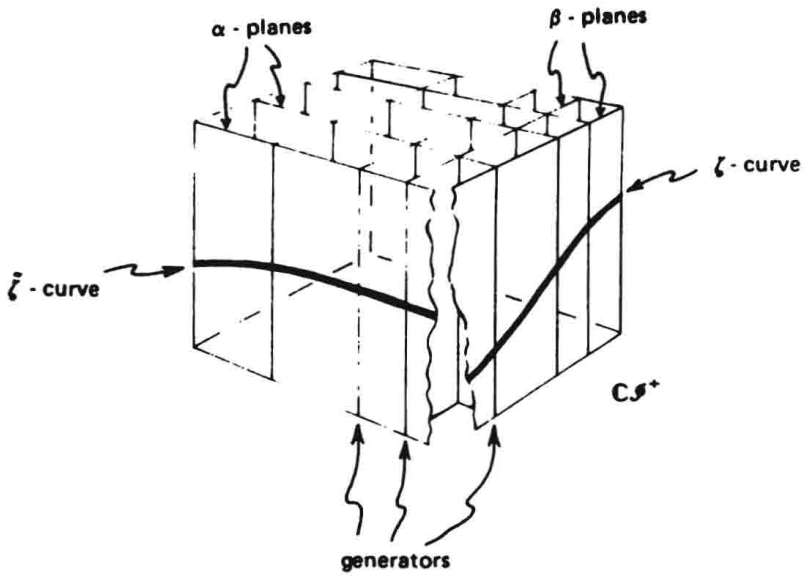


Fig. 8

field of M separately. A satisfactory twistor description of the combined fields must await further developments.

We have seen something of the remarkable way that twistor theory appears to have tied in, rather unexpectedly, with general relativity (albeit only partially, so far). So it seems not unreasonable to anticipate that such suitable "further developments" may indeed eventually take place. (For some recent suggestions in this direction see Reference 25.) And if they do, Einstein's theory should become fully describable in a natural way within the twistor program. In our opinion, it is imperative for the deeper aims of the program that Einstein's theory *can* indeed be so handled. The profound insights that general relativity has provided must be incorporated if twistor theory is to lay any claims in pointing to a more fundamental description of nature. And if they can be so incorporated, it seems likely these insights may also be helpful in leading to a twistor description of the other interactions of nature. Thanks to Einstein, the gravitational interaction is now understood in a marvelously detailed way. His theory provides the twistor formalism with a profound challenge and with many hopes for future new insights.

References

1. L. P. Hughston, *A Particle Classification Scheme Based on the Theory of Twistors*, Springer Lecture Notes in Physics, Springer, New York (1979).
2. R. Penrose and G. A. J. Sparling, Twistor theory of particles, *Rev. Mod. Phys.* (to appear).
3. M. Ko, M. Ludvigsen, E. T. Newman, and K. P. Tod, The theory of $\mathcal{H}$-space, preprint, Pittsburgh, 1978; *Phys. Rep. C.* (to appear).
4. J. A. Todd, *Projective and Analytical Geometry*, Pitman, London (1947).
5. R. Penrose, *J. Math. Phys.* **8**, 345–366 (1967).

6. R. Penrose, Relativistic symmetry groups, in *Group Theory in Non-Linear Problems*, A. O. Barut (ed.), Reidel, Dordrecht (1974), pp. 1–58.
7. R. Penrose and M. A. H. MacCallum, *Phys. Rep.* **6C**, 241–316 (1972).
8. R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics, and All That*, Benjamin, New York (1964).
9. R. Penrose, Twistor theory, its aims and achievements, in *Quantum Gravity*, C. J. Isham, R. Penrose, and D. W. Sciama (eds.), Clarendon, Oxford (1975), pp. 268–407.
10. R. Penrose, *J. Math. Phys.* **10**, 38–39 (1969).
11. R. C. Gunning and H. Rossi, *Analytic Functions of Several Complex Variables*, Prentice-Hall, Englewood Cliffs, New Jersey (1965).
12. R. Penrose, *Techniques of Differential Topology in Relativity*, SIAM, Philadelphia (1972).
13. R. Penrose, Structure of space-time, in *Battelle Rencontres*, C. M. DeWitt and J. A. Wheeler (eds.), Benjamin, New York (1968), pp. 121–235.
14. M. J. Crampin and F. A. E. Pirani, Twistors, symplectic structure and Lagrange's identity, in *Relativity and Gravitation*, Ch. G. Kuper and A. Peres (eds.), Gordon and Breach, London (1971), pp. 105–110.
15. S. S. Chern and J. K. Moser, *Acta Math.* **133**, 219–271 (1974).
16. R. Penrose, *Gen. Rel. Grav.* **7**, 31–53 (1976).
17. J. Morrow and K. Kodaira, *Complex manifolds*, Holt, Rinehart and Winston, New York (1971).
18. K. Kodaira, *Am. J. Math.* **85**, 79–94 (1963).
19. R. O. Hansen, E. T. Newman, R. Penrose, and K. P. Tod, *Proc. R. Soc. London Ser. A* **363**, 445–468 (1978).
20. R. S. Ward, *Proc. R. Soc. London Ser. A* **363**, 289–295 (1978).
21. W. D. Curtis, D. E. Lerner, and F. R. Miller, *J. Math. Phys.* **19**, 2024–2027 (1978).
22. S. W. Hawking, *Phys. Lett.* **60A**, 81–83 (1977).
23. M. Ko, E. T. Newman, and R. Penrose, *J. Math. Phys.* **18**, 58–64 (1977).
24. H. Komatsu (ed.), *Hyperfunctions and pseudo-differential equations*, Springer Lecture Notes in Mathematics, No. 287, Springer, Berlin (1973).
25. R. Penrose, in *Proceedings of the 1978 Kansas Conference on Complex Manifold Techniques in Physics*, D. E. Lerner and P. Sommers (eds.), Pitman, New York (1979).

Spaces and Manifolds Defined in This Chapter

CM	compactified complex Minkowski, 287	$\mathbf{P}\mathcal{N}(\mathcal{S})$	unscaled null geodesics, with respect to $\mathcal{S}$, 321
$\mathbf{CM}^I$	complexified Minkowski, 285	**PN**	null projective twistors, 291
$\mathbf{CP}^n$	complex projective, 286	**PN***	null dual projective twistors, 291
$\mathcal{H}(\mathcal{S})$	complex space-time associated with $\mathbf{P}\mathcal{T}(\mathcal{S})$, 325	$\mathbf{P}\mathcal{T}$	α surfaces in M, 313
M	complex space-time, 312	$\mathbf{P}\mathcal{T}(\mathcal{S})$	projective hypersurface twistors, 322
M	compactified real space-time, 288	**PT**	projective twistor, 286
$\mathbf{M}^I$	real Minkowski, 285	**PT***	projective dual twistors, 288
N	null twistors, 290	**T**	twistors, 285
$\mathbf{P}\mathcal{N}$	unscaled null geodesics, 309	**T***	dual twistors, 288
		$\mathcal{T}$	nonprojective twistors, 314
		$\mathcal{T}(\mathcal{S})$	hypersurface twistors, 322

114
NULL HYPERSURFACE INITIAL DATA FOR CLASSICAL FIELDS OF ARBITRARY SPIN AND FOR GENERAL RELATIVITY

Widely circulated preprint, *Cornell* (1961). Also circulated in P. G. Bergmann's ARL Technical Documentary Report (1963) 63–56; *Gen. Relativity Gravitation* **12** (1980), no. 3, 225–64

Largely written at Cornell University in 1961, this article remained unpublished (though circulated in 1963) until it appeared as a 'golden oldie' in *General Relativity and Gravitation* (GRG) in 1980. Two ingredients constituted the prime focus of this article, one being the notion of an *exact set of fields*, a concept originated in the article of Chapter 12, and the other, a formula which explicitly tells us how a massless free field is determined from its free null-data on an initial null hypersurface. This formula is called the 'generalized Kirchhoff–d'Adhémar formula' in *Spinors and Space-Time*, Volume 1, 1984, by R. Penrose and W. Rindler, Cambridge University Press; see also Chapter 35. Concepts of intrinsic geometry for a null hypersurface are related to those of Chapter 55.

576 Null hypersurface initial data for classical fields of arbitrary spin and for general relativity. Widely circulated preprint, *Cornell* (1961). Also circulated in P. G. Bergmann's ARL Technical Documentary Report (1963) 63–56; *Gen. Relativity Gravitation* **12** (1980), no. 3, 225–64

Null Hypersurface Initial Data for Classical Fields of Arbitrary Spin and for General Relativity[1,2]

R. PENROSE

Department of Astronomy, Cornell University, Ithaca, New York[3]

Abstract

A form of initial value problem is considered in which the initial hypersurface is not space-like but null. This approach has the striking advantage over the more usual Cauchy problem that all constraints (initial data equations) are eliminated from the theory, for a wide class of interacting fields in special relativity and also for general relativity. The theory is most naturally described in terms of the two-component spinor calculus, for which an elementary introduction is given here. A general scheme for interacting fields, which holds both in special and general relativity, is presented which describes all fields in terms of sets of irreducible spinors. The concept of an exact set of such spinors is introduced and it is shown that this concept is the appropriate one for an initial value problem on a null cone without constraints. The initial data can be expressed in the form of a complex number, called a null datum, defined at each point of the null cone, one corresponding to each spinor. There is the curious feature of these null data that apparently it is sufficient here, to have one-half as much information per point as in the corresponding Cauchy problem. The classical Maxwell–Dirac theory and the Einstein–Maxwell theory are two examples that can be put into the form of exact sets. The Einstein empty-space equations are also of particular note, and in this case the null datum describes essentially the intrinsic geometry of the null cone. The argument given here as applied to a general exact set is incomplete in two important respects. Firstly it depends on the null data being analytic, and secondly the initial hypersurface must be a cone. However, both these restrictions are removed in the case of certain elementary fields called basic free fields, examples of which are the Weyl neutrino field, the free Maxwell field, and the linearized gravitational field. For these cases a simple explicit formula is introduced which expresses the field at any point in terms of the null datum, as an integral taken over the intersection of the initial null hypersurface with the null cone of the point.

[1] This article originally appeared in 1963 in Aerospace Research Laboratories 63-56 (P.G. Bergmann). It is an important and oft-cited work, but as it has never been published in a widely distributed journal, it is generally inaccessable to the relativity community. This regrettable situation is hereby rectified–Ed.

[2] This work was done while the author was at Princeton, Syracuse, and Cornell Universities, visiting under a NATO Fellowship administered by the Department of Scientific and Industrial Research in London. The work at Syracuse was supported by the Aeronautical Research Laboratory and at Cornell by the National Science Foundation.

[3] Present address: Mathematical Institute, University of Oxford, Oxford, England.

Null hypersurface initial data for classical fields of arbitrary spin and for general relativity. Widely circulated preprint, *Cornell* (1961). Also circulated in P. G. Bergmann's ARL Technical Documentary Report (1963) 63–56; *Gen. Relativity Gravitation* **12** (1980), no. 3, 225–64

§(1): *Introduction*

It has generally been customary, in relativity theory, to work largely in terms of spacelike hypersurfaces and timelike directions. This is natural, for the older Newtonian idea of a three-dimensional spatial universe developing with time seems close to our experiences and is easy to visualize. In this paper, however, I wish to stress another point of view: that null hypersurfaces and null directions are, in fact, much more convenient and may perhaps also be regarded as more fundamental.

There are some good physical indications of this. For the two fields most important to relativity theory, namely, electromagnetism and gravitation, having zero rest mass, propagate in null directions and along null hypersurfaces. But there is also a strong mathematical reason for considering null directions to be more elementary than timelike ones. This is essentially that the (two-component) spinor calculus is more elementary than the tensor calculus. Spinors represent a breakdown of tensors into simpler parts, and the simplest spinors (other than scalars), namely, the one-index two-component spinors, define *null* directions in space-time.

It will turn out here that by a combination of a consistent use of spinors, and the use of null hypersurfaces on which to specify the initial data, some remarkable simplifications and unifications can be obtained which do not occur in the usual Cauchy approach [1, 2]. The most important of these simplifications appears to be the complete *elimination of all constraints* from the theory. The essential initial data can be given in the form of one complex number per point of the hypersurface for each elementary field, the specification of these complex numbers being essentially unrestricted. This apparently holds both for special and for general relativity[4] with a wide choice of possible interacting fields.[5] A further feature of this approach is the existence of very simple integrals for the basic free fields of the theory.[6] Many interacting fields can also be treated by an extension of this method involving Feynman-like graphs, but these will be left to a later paper.

Since the techniques employed in this paper will involve two-component spinor techniques almost exclusively, I shall first review the essential basic ideas, notations, and geometrical representations that will be required. For other more extensive accounts see, for example [6, 12; also 13, 14].

Greek letters will be used for tensor indices and run over 0, 1, 2, 3. The

[4] Some preliminary results of a similar nature in general relativity have also been obtained by Bondi and Sachs [3].

[5] The importance to cosmology of having an initial value problem that can be stated in terms of the null cone of an observer has been stressed to me particularly by D. W. Sciama and E. Schuecking. This fact that *all* the direct large-scale observations of the universe that have been made lie on what is essentially a null cone, has also been emphasized by J. L. Synge and others.

[6] Some special cases of these integrals are already known [4, 5].

metric tensor $g_{\mu\nu}$ (and its inverse $g^{\mu\nu}$) have signature −2 and can therefore be reduced to the form diag $\{1, -1, -1, -1\}$ at any one point. In special relativity, this can of course also be done globally. Spinor indices are of two types and will be denoted by capital Roman letters, unprimed or primed.[7] They run over the two values 0, 1 or 0′, 1′. Tensor indices are translated into spinor form by means of two inverse quantities $\sigma_\mu{}^{AB'}$, $\sigma^\mu{}_{AB'}$ as follows:

$$X^{AB'CD'}{}_{EF'} = \sigma_\lambda{}^{AB'}\sigma_\mu{}^{CD'}X^{\lambda\mu}{}_\nu\sigma^\nu{}_{EF'}$$

and inversely

$$X^{\lambda\mu}{}_\nu = \sigma^\lambda{}_{AB'}\sigma^\mu{}_{CD'}X^{AB'CD'}{}_{EF'}\sigma_\nu{}^{EF'}$$

so that each tensor index is represented by a pair of spinor indices, one unprimed and one primed. (The summation convention is used for tensor or spinor indices throughout this paper.) Unprimed indices may be contracted only with unprimed indices and primed indices only with primed indices. The $\sigma^\mu{}_{AB'}$ is a (2 × 2) Hermitian matrix for each value of μ, and satisfies

$$g_{\mu\nu}\sigma^\mu{}_{AB'}\sigma^\nu{}_{CD'} = \epsilon_{AC}\epsilon_{B'D'} \tag{1.1}$$

where ϵ_{AC}, $\epsilon_{B'D'}$ together with ϵ^{AC}, $\epsilon^{B'D'}$ are Levi-Civita symbols (skew-symmetric expressions with $\epsilon_{01} = \epsilon_{0'1'} = \epsilon^{01} = \epsilon^{0'1'} = 1$) and are used for raising and lowering spinor indices:

$$\xi^A = \epsilon^{AB}\xi_B, \quad \xi_B = \xi^A\epsilon_{AB}, \quad \eta^{A'} = \epsilon^{A'B'}\eta_{B'}, \quad \eta_{B'} = \eta^{A'}\epsilon_{A'B'} \tag{1.2}$$

(Note the ordering of indices.) In particular, it follows from (1.1) and $\sigma^\mu{}_{AB'}\sigma_\nu{}^{AB'} = \delta^\mu_\nu$ that

$$\sigma_\mu{}^{AB'} = \epsilon^{AC}\epsilon^{B'D'}\sigma^\nu{}_{CD'}g_{\mu\nu}$$

However, being skew-symmetric, the ϵ's do not really define a "metric" for spin space. It is important to observe, moreover, that the relation $\xi_A\eta^A = 0$ holds *if* and *only if* one of ξ_A, η_A is a scalar multiple of the other.

In special relativity we may choose $g_{\mu\nu}$ = diag $\{1, -1, -1, -1\}$ everywhere. Then a convenient representation for $2^{1/2}\sigma^\mu{}_{AB'}$ everywhere, which satisfies (1.1), is the unit matrix together with the Pauli matrices. The relation between a vector x_μ and its spinor equivalent $x_{AB'}$ is then given by

$$\begin{pmatrix} x_{00'} & x_{01'} \\ x_{10'} & x_{11'} \end{pmatrix} = 2^{-1/2}\begin{pmatrix} x_0 + x_1 & x_2 - ix_3 \\ x_2 + ix_3 & x_0 - x_1 \end{pmatrix} \tag{1.3}$$

$$(x_0, x_1, x_2, x_3) = 2^{-1/2}(x_{00'} + x_{11'}, x_{00'} - x_{11'}, x_{01'} + x_{10'}, ix_{01'} - ix_{10'})$$

Spinors are naturally *complex* objects, the notion of reality having no invariant meaning for them (except in the special case of scalars). However,

[7] For typographical reasons, primed indices will be used here rather than the more usual dotted indices. This agrees with the notation of [13, 14].

reality of a tensor is interpreted as a *Hermitian* property of the corresponding spinor, the complex conjugate of an unprimed index being a primed index and vice versa. For example, the complex conjugate[8] of $X^{AB'CD'}{}_{EF'}$ is $\bar{X}^{A'BC'D}{}_{E'F}$ whence the condition that $X^{\lambda\mu}{}_{\nu}$ be real is the Hermitian relation

$$X^{AB'CD'}{}_{EF'} = \bar{X}^{B'AD'C}{}_{F'E}$$

Now let us consider the geometrical interpretation of a one-index spinor ξ_A. The vector u_μ whose spinor equivalent is given by

$$u_{AB'} = \xi_A \bar{\xi}_{B'} \tag{1.4}$$

is real and null, for, by (1.1), (1.2), tensor contractions translate as spinor contractions and so $u_\mu u^\mu = u_{AB'} u^{AB'} = |\xi_A \xi^A|^2 = 0$. Thus, any one-index spinor determines a *null vector.* Also, conversely, it is not hard to see that any null vector that points into the future can be expressed in the form (1.4). However, ξ_A contains more information than u_μ since a phase multiplier for ξ_A does not affect (1.4). The phase of ξ_A also has a geometrical interpretation. The bivector $p_{\mu\nu}$ with spinor equivalent

$$p_{AB'CD'} = \xi_A \xi_C \epsilon_{B'D'} + \epsilon_{AC} \bar{\xi}_{B'} \bar{\xi}_{D'} \tag{1.5}$$

has the form

$$p_{\mu\nu} = u_\mu w_\nu - u_\nu w_\mu \tag{1.6}$$

where

$$w_{AB'} = \xi_A \bar{\eta}_{B'} + \eta_A \bar{\xi}_{B'}, \qquad \xi_A \eta^A = 1$$

The vector w_μ is real and spacelike (with $w_\mu w^\mu = -2$), orthogonal to u_μ, and is determined by ξ_A up to the addition of a multiple of u_μ. Thus ξ_A determines, not only the null vector u_μ, but also a certain null half-plane element containing the null direction, namely, the half-plane of directions of $w_\mu + \lambda u_\mu$. A spinor may therefore be conveniently pictured as a kind of "flag" (see Figure 1.). The flagpole represents the null vector u_μ and the flag itself the halfplane element determined by u_μ and w_μ. This flag and flagpole give a complete geometrical realization of ξ_A apart from a sign [(1.4) and (1.5) are invariant under $\xi_A \to -\xi_A$]. This sign ambiguity is, however, essential for such a direct geometrical interpretation, since a continuous spacial rotation through an angle 2π changes ξ_A continuously into $-\xi_A$.

Note that the null *direction* of the flagpole is unchanged if ξ_A is multiplied by any nonzero complex scalar λ. The *extent*[9] of the flagpole is, however, multiplied by a factor $|\lambda|^2$ and the flag rotates through an angle 2 arg λ. *Two* complete rotations of the flag thus are necessary to correspond to a phase change

[8]The use here of the bar for complex conjugates of spinors is made for reasons of clarity and of consistency with the notation for scalars, as in [13, 14].

 Null hypersurface initial data for classical fields of arbitrary spin and for general relativity. Widely circulated preprint, *Cornell* (1961). Also circulated in P. G. Bergmann's ARL Technical Documentary Report (1963) 63–56; *Gen. Relativity Gravitation* **12** (1980), no. 3, 225–64

Fig. 1. A spinor ξ_A is represented geometrically, up to a sign, as a "flag." The flagpole is a null vector u_μ and the flag itself is a half-plane element spanned by u_μ and a spacelike vector w_μ orthogonal to it. To distinguish the representation from that of a primed spinor $\zeta_{B'}$, a screw sense can be associated with the flagpole.

through an angle 2π for ξ_A. The representation of a spinor $\zeta_{B'}$ with a single primed index is essentially the same as that for ξ_A except that the flag rotates in the opposite direction when $\zeta_{B'}$ is multiplied by λ. It is probably significant, therefore, to associate a certain *screw sense* with a spinor ξ_A and to associate the opposite screw sense with $\zeta_{B'}$. This is related to the spin properties of zero restmass particles.

It is conceptually very useful that there is a clear-cut geometrical representation of spinors, for it removes most of the abstractness that is generally associated with "spin space." In fact, it is *only* the sign ambiguity that prevents spinors from being just as good geometrical objects as tensors. This sign ambiguity does not seem to cause any particular trouble, however, and it is not present for spinors having an even number of indices.

Since spinors have a geometrical interpretation, the notation of covariant derivative has a natural unambiguous definition in terms of parallel displacement. In special relativity we may choose a Minkowski form for the metric tensor and constant components for $\sigma^\mu{}_{AB'}$ [for example, those corresponding to (1.3)]. Then the operator ∂_μ (or $\partial_{AB'}$) of covariant differentiation is simply the coordinate derivative operator $\partial/\partial x^\mu$ (or $\partial/\partial x^{AB'}$). In general relativity, or in special relativity with curvilinear coordinates, covariant differentiation is distinct from coordinate differentiation but I shall still use the symbol ∂_μ, or

[9]The term "extent" will be used here for null vectors in place of "length" since the length of a null vector is necessarily zero. The extent of a null vector is not to be thought of as an absolute numerical quantity in itself, but if two null vectors have the same direction the *ratio* of their extents will be a numerical quantity (namely, the ratio of the null vectors). Similarly any null geodesic will have an extent that can be expressed numerically only in terms of the extent of, say, one of its tangent vectors. Extents along a null geodesic can be compared, since parallel transport carries tangent null vectors into tangent null vectors.

more usually

$$\partial_{AB'} \equiv \sigma^{\mu}{}_{AB'}\partial_{\mu}$$

to denote *covariant* differentiation[10] to stress the fact that it is this operator, rather than $\partial/\partial x^{AB'}$, that has physical meaning. In general relativity the $\partial_{AB'}$ operators do not commute, however (see, for example [15]). The covariant derivatives of $g_{\mu\nu}$, $\sigma^{\mu}{}_{AB'}$, ϵ_{AC}, $\epsilon_{B'D'}$ are all required to be zero. This fixes the definitions of the spinor affine connections, $\Gamma^{A}_{B\mu}$, $\Gamma^{A'}_{B'\mu}$ in terms of coordinate derivatives of $\sigma^{\mu}{}_{AB'}$, but I shall not be concerned with the details of such coordinate-dependent quantities here, the formal properties of $\partial_{AB'}$ being all that will be required [13].

The quantities $\sigma^{\mu}{}_{AB'}$ should be regarded as merely a dictionary for translating from tensor notation into spinor notation. They do not have any geometrical significance unless it is to be regarded as significant when a spinor index takes on a particular value. Then $\sigma^{\mu}{}_{AB'}$ may be regarded as determining a "null tetrad":

$$l^{\mu} = \sigma^{\mu}{}_{00'}, \quad n^{\mu} = \sigma^{\mu}{}_{11'}, \quad m^{\mu} = \sigma^{\mu}{}_{01'}, \quad \bar{m}^{\mu} = \sigma^{\mu}{}_{10'} \tag{1.7}$$

where l^{μ}, n^{μ} are null and real, and m^{μ} is null and complex (its real and imaginary parts are orthogonal and spacelike) and orthogonal to n^{μ}, l^{μ} [see (1.1)]. However, (1.7) is a little misleading because the expressions are not vectors as they stand and behave differently with respect to covariant differentiation. This point is discussed more fully in [14].

The full elegance of the spinor calculus is really only brought out when *all* indices are translated into a two-component spinor form. Expressions with skew-symmetric tensor indices have particularly simple spinor analogs. The spinor equivalent of the Maxwell field tensor $F_{\mu\nu}$ can be written

$$F_{AC'BD'} = \phi_{AB}\epsilon_{C'D'} + \epsilon_{AB}\bar{\phi}_{C'D'}$$

where ϕ_{AB} is symmetric. The spinor ϕ_{AB} then conveniently represents the electromagnetic field, the Maxwell free-field equations being

$$\partial^{AC'}\phi_{AB} = 0 \tag{1.8}$$

Similarly, if the Riemann curvature tensor $R_{\mu\nu\rho\sigma}$ satisfies the Einstein empty space condition $R^{\mu}{}_{\nu\mu\sigma} = 0$, its spinor equivalent has the form

$$R_{AE'BF'CG'DH'} = \psi_{ABCD}\epsilon_{E'F'}\epsilon_{G'H'} + \epsilon_{AB}\epsilon_{CD}\bar{\psi}_{E'F'G'H'} \tag{1.9}$$

where ψ_{ABCD} is totally symmetric [16, 13]. The Bianchi identities become

$$\partial^{AE'}\psi_{ABCD} = 0 \tag{1.10}$$

If the Ricci tensor $R^{\mu}{}_{\nu\mu\sigma}$ does *not* vanish, then the Riemann tensor cannot be represented simply in this way [see (3.28)]. However, Weyl's conformal curva-

[10] This, again, agrees with [13].

ture tensor $C_{\mu\nu\rho\sigma}$, which differs from the Riemann tensor by an expression constructable from the Ricci tensor, has the same symmetries as the Riemann tensor and satisfies $C^{\mu}{}_{\nu\mu\sigma} = 0$. Thus, $C_{\mu\nu\rho\sigma}$ may be used in place of $R_{\mu\nu\rho\sigma}$ in (1.9) and a symmetric ψ_{ABCD} is still defined.[11] This spinor may then be thought of as representing the *purely gravitational* part of the curvature, the parts due to the energy tensor of the matter fields having been subtracted out. These matter fields then appear in a right-hand side to equation (1.10); that is, they give a source term for the ψ_{ABCD} field.

It is interesting to compare (1.8) and (1.10) with the Weyl equation for the neutrino:

$$\partial^{AB'}\nu_A = 0.$$

These three equations are particular cases of the basic zero rest-mass equation for spin $\frac{1}{2}n$ (>0):

$$\partial^{AZ'}\theta_{AB\cdots N} = 0 \tag{1.11}$$

where $\theta_{AB\cdots N}$ has n symmetric unprimed indices.

In special relativity we have

$$\partial_{AY'}\partial^{AZ'} \equiv \tfrac{1}{2}\delta^{Z'}_{Y'}\,\Box \tag{1.12}$$

where $\Box$ is the D'Alembertian operator $\partial_{AB'}\partial^{AB'}$. Applied to (1.11), this gives D'Alembert's equation for each component of $\theta_{AB\cdots N}$. This does not hold in general relativity, however [13].

In two-component spinor notation the free field Dirac equation takes the form [6, 9]

$$\partial^{AB'}\alpha_A = \frac{m}{2^{1/2}}\beta^{B'}, \qquad \partial_{AB'}\beta^{B'} = -\frac{m}{2^{1/2}}\alpha_A \tag{1.13}$$

so it may be regarded as a pair of coupled zero rest-mass fields α_A and $\beta^{B'}$ each of which is a source for the other. The more usual Dirac four-component spinor is essentially the direct sum of the spinors α_A, $\beta^{B'}$. A four-component spinor index may therefore be regarded as a direct sum of two two-component spinor indices, one unprimed and one primed.

Equation (1.13) may be generalized to higher spin [17]:

$$\begin{aligned}\partial^{AP'}\theta^{Q'R'\cdots T'}_{ABC\cdots G} &= \frac{m}{2^{1/2}}\phi^{P'Q'R'\cdots T'}_{BC\cdots G}\\ \partial_{AP'}\phi^{P'Q'R'\cdots T'}_{BC\cdots G} &= -\frac{m}{2^{1/2}}\theta^{Q'R'\cdots T'}_{ABC\cdots G}\end{aligned} \tag{1.14}$$

where $\theta\,\vdots\,$ and $\phi\,\vdots\,$ are symmetric in all primed indices and also in all unprimed indices. The spin is half the total number of indices of $\theta\,\vdots\,$ (or $\phi\,\vdots\,$). Equations

[11] The definition of ψ_{ABCD} given here agrees with that of [14] but is one-half that given in [13], assuming the Riemann tensor to have the sign of that given in [13].

(1.14) may also be considered as a pair of interacting fields (but see the end of the next section).

§(2): *Null-Cone Initial Data*

Consider a set of interacting fields:

$$\theta_{AB\cdots D}^{P'Q'\cdots T'}, \quad \phi_{A\cdots F}^{P'\cdots R'}, \ldots, \quad \chi_{A\cdots G}^{P'\cdots U'} \tag{2.1}$$

where each expression is totally symmetric in all primed indices and also in all unprimed indices.[12] (Either or both sets of indices may be vacuous.) The *field equations* will be supposed to be a set of covariant differential relations connecting the fields. There may also be some relations relating commutators of $\partial_{AB'}$'s to expressions involving the fields. Of course, in special relativity these commutators vanish.

Such a set of fields and field equations, etc. will be called *exact* if they have the following two properties.

(i) At any event O, all the symmetrized (covariant) derivatives

$$\partial_{(K}{}^{(W'}\partial_{L}{}^{X'}\cdots\partial_{N}{}^{Z'}\theta_{AB\cdots D)}^{P'Q'\cdots T')}, \ldots, \partial_{(K}{}^{(W'}\cdots\partial_{M}{}^{Y'}\chi_{A\cdots G)}^{P'\cdots U')} \tag{2.2}$$

[including the "zero times" differentiated expressions (2.1)] are *independent*. That is, they can take independently arbitrary values at O. (The brackets around indices denote complete symmetrization.)

(ii) At any event O, all the unsymmetrized derivatives

$$\partial_{K}{}^{W'}\cdots\partial_{M}{}^{Y'}\theta_{A\cdots D}^{P'\cdots T'}, \ldots, \partial_{K}{}^{W'}\cdots\partial_{M}{}^{Y'}\chi_{A\cdots G}^{P'\cdots U'}$$

of the expressions (2.1) are completely determined by the symmetrized derivatives (2.2) [and (2.1)].

It appears that the most important known fields of physics, including general relativity, can all be put into the form of exact sets of interacting fields.

As one example, consider the Maxwell free-field equations in flat space: $\partial^{AS'}\phi_{AB} = 0$. This states simply that $\partial_{G}{}^{S'}\phi_{AB}$ is already symmetric in GAB. Next consider $\partial_{F}{}^{R'}\partial_{G}{}^{S'}\phi_{AB}$. This is symmetric in GAB and, since the ∂'s may be commuted, also symmetric in FAB. It is therefore symmetric in $FGAB$ and hence, again because the ∂'s commute, also symmetric in R', S'. Continuing, we obtain the result that

$$\partial_{D}{}^{P'}\cdots\partial_{G}{}^{S'}\phi_{AB} = \partial_{(D}{}^{(P'}\cdots\partial_{G}{}^{S')}\phi_{AB)} \tag{2.3}$$

so that condition (ii) for exactness is trivially satisfied. Condition (i) is satisfied too, because of the linearity of the field equations and the fact that no further symmetries can be imposed on the expressions (2.3).

The free-field Dirac equations (1.13) for $\alpha_{A}, \beta^{B'}$ in flat space also give an

[12] Each is therefore irreducible under the group of homogeneous Lorentz transformations.

 Null hypersurface initial data for classical fields of arbitrary spin and for general relativity. Widely circulated preprint, *Cornell* (1961). Also circulated in P. G. Bergmann's ARL Technical Documentary Report (1963) 63–56; *Gen. Relativity Gravitation* **12** (1980), no. 3, 225–64

exact set. For example,

$$\partial_C{}^{B'}\alpha_A = \partial_{(C}{}^{B'}\alpha_{A)} - \frac{m}{2(2^{1/2})}\epsilon_{CA}\beta^{B'}$$

and it is not hard to see how this continues for higher derivatives. (The skew parts of any spinor expression are always expressible in terms of contractions.) The case of the classical Maxwell–Dirac coupled equations is a little more involved, since here a potential $\theta_{AB'}$ for the ϕ field is needed. If the Lorenz condition is adopted, this gives

$$\partial_{BP'}\theta_A{}^{P'} = \phi_{AB}, \qquad \partial^{AQ'}\theta_A{}^{P'} = -\bar{\phi}^{P'Q'} \tag{2.4}$$

and the Maxwell–Dirac equations are

$$\partial^{AP'}\alpha_A = ie\theta^{AP'}\alpha_A + \frac{m}{2^{1/2}}\beta^{P'}, \qquad \partial_{AP'}\beta^{P'} = ie\theta_{AP'}\beta^{P'} - \frac{m}{2^{1/2}}\alpha_A \tag{2.5}$$

$$\partial^{AP'}\phi_{AB} = k\ \{\alpha_B\bar{\alpha}^{P'} + \beta_B\bar{\beta}^{P'}\} \tag{2.6}$$

(e, m, and k being suitable constants). The fields $\theta_A{}^{P'}$, ϕ_{AB}, α_A, $\beta^{P'}$ can, without great difficulty, be shown to give an exact set.

In general relativity the situation is more complicated because of the non-commutativity of the ∂'s. However, it has been shown [13, p. 199], for the Einstein–Maxwell equations, ϕ_{AB} and ψ_{ABCD} [see (1.9) *et seq.* and (3.28)] given an exact set, where,

$$\partial^{AP'}\phi_{AB} = 0, \qquad \partial^{AP'}\psi_{ABCD} = -k\bar{\phi}^{P'Q'}\partial_{BQ'}\phi_{CD} \tag{2.7}$$

k being a positive constant and where there are also the commutator relations generated by

$$\{\partial_A{}^{P'}\partial_B{}^{Q'} - \partial_B{}^{Q'}\partial_A{}^{P'}\}\,\xi_C = \epsilon^{P'Q'}\psi_{ABCD}\xi^D + k\epsilon_{AB}\bar{\phi}^{P'Q'}\phi_{CD}\xi^D \tag{2.8}$$

for any ξ_C. Of course the Einstein empty-space equations are a special case of this with $\phi_{AB} = 0$. Then ψ_{ABCD} gives an exact set. There are also many other sources for the gravitational field which, with ψ_{ABCD}, give an exact set. The presence of a λ term in Einstein's equations does not affect this.

Let us now examine, in the general case, the geometrical significance of the symmetrized derivatives (2.2). Suppose we are given a particular solution $\theta_{A\cdots D}^{P'\cdots T'}, \phi_{A\cdots F}^{P'\cdots R'}, \ldots, \chi_{A\cdots G}^{P'\cdots U'}$ of the field and commutator equations which is analytic[13] and for which the space-time is analytic and Riemannian[14] (possibly flat) in the relevant region. (The commutator relations for $\partial_{AB'}$ determine the geometry of the space-time.) Then, the field $\theta(X)_{A\cdots D}^{P'\cdots T'}$ at any event X

[13] Although the fields are complex, there is no suggestion intended that the fields be complex-analytic in some strong sense. The real and imaginary parts of the fields are simply analytic functions of four real variables.

[14] It seems probable that the theory can also be generalized to include spaces with torsion, see [15].

not "too far" from the event O can be expressed as a power series:

$$\theta(X)_{A\cdots D}^{P'\cdots T'} \overset{XO}{=} \theta(O)_{A\cdots D}^{P'\cdots T'} + x^{\mu}\partial_{\mu}\theta(O)_{A\cdots D}^{P'\cdots T'} + \tfrac{1}{2}x^{\mu}x^{\nu}\partial_{\mu}\partial_{\nu}\theta(O)_{A\cdots D}^{P'\cdots T'} + \cdots \tag{29}$$

The spinors on the right-hand side are to be evaluated at O after differentiation. The vector x^{μ} gives the *normal coordinates* of the point X relative to O, that is to say x^{μ} has the same direction at O as the geodesic OX and also the same length (or extent, if it is null). In Minkowski space-time, x^{μ} is simply the position vector of X relative to O. The spinors at the two points are to be compared, in the general case, by parallel transport along the geodesic OX and this is the meaning of the symbol $\overset{XO}{=}$ used in (2.9), (2.11). [One convenient way of doing this is to choose the $\sigma_{\mu}{}^{AB'}$ so that the tetrad defined by (1.7) is parallelly transported along OX. Then components may be equated in (2.9) and in (2.11).]

Next, choose x^{μ} to be a null vector, so that

$$x^{AB'} = l\xi^{A}\bar{\xi}^{B'} \tag{2.10}$$

where l is a real number and choose ξ^{A} to be covariantly constant along $0X$. Substituting (2.10) in (2.9) and contracting with $\xi^{A}\cdots\xi^{D}\bar{\xi}_{P'}\cdots\bar{\xi}_{T'}$ we get

$$\begin{aligned}\xi^{A}\cdots\xi^{D}\bar{\xi}_{P'}\cdots\bar{\xi}_{T'}\theta(X)_{A\cdots D}^{P'\cdots T'} &\overset{XO}{=} \xi^{A}\cdots\xi^{D}\bar{\xi}_{P'}\cdots\bar{\xi}_{T'}\theta(O)_{A\cdots D}^{P'\cdots T'} \\ &- l\xi^{A}\cdots\xi^{D}\xi^{E}\bar{\xi}_{P'}\cdots\bar{\xi}_{T'}\bar{\xi}_{U'}\partial_{(E}{}^{(U'}\theta(O)_{A\cdots D)}^{P'\cdots T')} \\ &+ \frac{l^2}{2!}\xi^{A}\cdots\xi^{D}\xi^{E}\xi^{F}\bar{\xi}_{P'}\cdots\bar{\xi}_{T'}\bar{\xi}_{U'}\bar{\xi}_{V'}\partial_{(E}{}^{(U'}\partial_{F}{}^{V'}\theta(O)_{A\cdots D)}^{P'\cdots T')} - \cdots\end{aligned} \tag{2.11}$$

The coefficients on the right-hand side are the *symmetrized* derivatives because of the symmetry of $\xi^{A}\cdots\xi^{G}\bar{\xi}_{P'}\cdots\bar{\xi}_{W'}$.

Define

$$\theta = \theta(\xi_{A};X) = \xi^{A}\cdots\xi^{D}\bar{\xi}_{P'}\cdots\bar{\xi}_{T'}\theta(X)_{A\cdots D}^{P'\cdots T'} \tag{2.12}$$

Equation (2.11) then tells us that knowledge of this complex function is equivalent to knowledge of all the symmetrized derivatives of $\theta_{A\cdots D}^{P'\cdots T'}$ evaluated at O. The points X, at which θ is defined, lie on the *null cone* $\mathfrak{N}$ of O (i.e., the points of all null geodesics through O). Choosing a particular ξ^{A} for each null direction at O, we have essentially just a complex number $\theta(X)$ defined for each point X of $\mathfrak{N}$. If ξ^{A} is replaced by $\lambda\xi^{A}$, then θ becomes $\lambda^{q}\bar{\lambda}^{p}\theta$, where $\theta_{A\cdots D}^{P'\cdots T'}$ has p primed and q unprimed indices.

This complex function θ of the points of $\mathfrak{N}$ will be called the *null datum* on $\mathfrak{N}$ for the field $\theta_{A\cdots D}^{P'\cdots T'}$. We have just seen that, in the analytic case, knowledge of the null datum on θ for $\theta_{A\cdots D}^{P'\cdots T'}$ is equivalent to knowledge of all the symmetrized derivatives of $\theta_{A\cdots D}^{P'\cdots T'}$ at O. Also, it is clear from (2.9) that knowledge of all the unsymmetrized derivatives at O gives us the whole field. The same holds for the other fields $\phi_{A\cdots F}^{P'\cdots R'}, \ldots, \chi_{A\cdots G}^{P'\cdots U'}$. Thus we see that condition (ii)

for exactness on the fields $\theta_{A\ldots D}^{P'\ldots T'}, \ldots, \chi_{A\ldots G}^{P'\ldots U'}$ tells us that the null data $\theta, \ldots, \chi$ on $\mathfrak{N}$ for these fields determine the fields uniquely throughout spacetime. Also, condition (i) for exactness tells us that these null data can be given *independently*, that is, there are *no constraints* (initial data equations).

The above argument has been given for the analytic case only. But it would appear that the result is true generally, without any such restrictions of analyticity. Any nonanalytic solution can be approximated by means of analytic solutions. However, this does not seem to lead to a general proof of the required result in the nonanalytic cases. There are important questions that arise here that are obscured in considerations of only analytic cases. For example, we expect the null data on the future half-cone $\mathfrak{N}^+$ to determine the fields in its interior (in a way which is "reasonable"), and correspondingly for the past half-cone $\mathfrak{N}^-$. We expect the fields in the region exterior to both $\mathfrak{N}^+$ and $\mathfrak{N}^-$ to depend on null data on both half-cones but in an "unreasonable" way, that is, regions of both $\mathfrak{N}^+$ and $\mathfrak{N}^-$ extending all the way to infinity are involved so that the local values of the fields depend on global properties of the null data. There is also the important question of stability [18]. Some of these matters will become clearer from an examination of the integrals of Section 4.

A curious feature of null data is the fact that one-half of the amount of information that is required in the usual Cauchy problem is apparently sufficient here. For example, the null datum $\phi_{AB}\xi^A\xi^B$ for the free Maxwell field gives us two real numbers (one complex number) per point of $\mathfrak{N}$, while for a spacelike hypersurface we require four real numbers per point (six for E, H minus two for the constraints $\operatorname{div} E = \operatorname{div} H = 0$). Similarly, for the gravitational field in empty space, the null datum is $\psi_{ABCD}\xi^A\xi^B\xi^C\xi^D$ (exactly how this is to be specified and its relation to the intrinsic geometry of $\mathfrak{N}$ will be discussed in more detail in the next section). This again gives two real numbers per point of $\mathfrak{N}$ (and there are no constraints) whereas for a spacelike hypersurface we require four (many constraints having been taken into account). This apparent disagreement seems to be related to the fact that $\mathfrak{N}$ consists of two pieces, $\mathfrak{N}^+$ and $\mathfrak{N}^-$. In the exterior region, the determination of the fields by the null data perhaps more closely resembles a two-hypersurface (or sandwich) formalism than a Cauchy problem, half the required information lying on each hypersurface, one in the future and one in the past. In the region interior to $\mathfrak{N}^+$, say, the situation is different. Here, we expect the fields to depend only on the null data on $\mathfrak{N}^+$ so we can ignore the data on $\mathfrak{N}^-$. But the data on $\mathfrak{N}^+$ alone are in a sense only half sufficient in that they allow us to proceed in one direction in time only, namely, into the future.[15]

[15] This explanation leaves something to be desired, however. It is easy to produce an example of a *spacelike* hypersurface whose domain of dependence is precisely the same as that of $\mathfrak{N}^+$. There is probably a subtler reason for this halving of the number of initial data functions. See [5] and [18].

The null data for the Dirac field (1.13) are $\alpha_A \xi^A$, $\beta^{B'} \bar{\xi}_{B'}$. This gives two complex numbers per point instead of the usual four complex numbers per point for a spacelike hypersurface. In the case of the classical Maxwell-Dirac theory, in addition to the Maxwell null datum and Dirac null data, we also require the null datum $\theta_A{}^{B'} \xi^A \bar{\xi}_{B'}$ for the potential. However, this only affects the gauge and a convenient choice is to make this zero. Then the potential is determined uniquely.

For the Dirac higher-spin fields $\theta \dots$, $\phi \dots$ satisfying (1.14), in order to get an exact set it is necessary to introduce more fields by differentiation. If $\theta_{A \dots G}^{Q' \dots T'}$ had p primed and q unprimed indices, then we obtain p new fields by successive differentiation with contraction each time on a primed index. Similarly, we get q other fields (including $\phi \dots$ if the contraction takes place successively instead on unprimed indices. Each of these $(p + q + 1)$ fields is symmetric in all unprimed and all primed indices and together they give an exact set. Thus the null data are $p + q + 1 = 2s + 1$ complex numbers per point of $\mathfrak{N}$, where $s = \frac{1}{2}(p + q)$ is the spin, as expected.

§(3): *Null Hypersurfaces*

A general null hypersurface is a three-dimensional submanifold of space-time (usually possessing singularities) with the property that all its normal vectors are *null* vectors. Such a hypersurface $\mathfrak{N}$ then has the curious property that its normal directions are also *tangent* to $\mathfrak{N}$ (since null vectors are orthogonal to themselves). Furthermore these normal directions are the tangents to a system of null *geodesics* of the space-time, lying on $\mathfrak{N}$, one through each point of $\mathfrak{N}$. These null geodesics will be called the *generators* of $\mathfrak{N}$. Of course, in special relativity, these generators are straight lines.

To see that these null lines are geodesics (or straight), let the hypersurface $\mathfrak{N}$ be given by $u = 0$, where u is a function of position. Suppose that the neighboring hypersurfaces $u = \text{const}$ are also all chosen to be null. Then the normals $u_\mu = \partial_\mu u$ satisfy $u^\mu u_\mu = 0$, so differentiating we get

$$2u^\mu \partial_\nu u_\mu = 0 \tag{3.1}$$

but u_μ is a gradient field, whence its curl vanishes:

$$\partial_\mu u_\nu - \partial_\nu u_\mu = 0 \tag{3.2}$$

Contracting with u^μ and using (3.1) we get

$$(u^\mu \partial_\mu) u_\nu = 0 \tag{3.3}$$

From this we see that the direction of the vector u_ν at one point, and that of the corresponding vector at a neighboring point in the direction u_μ, are parallel. That is, the u_μ are tangents to geodesics in space-time [15].

 Null hypersurface initial data for classical fields of arbitrary spin and for general relativity. Widely circulated preprint, *Cornell* (1961). Also circulated in P. G. Bergmann's ARL Technical Documentary Report (1963) 63–56; *Gen. Relativity Gravitation* **12** (1980), no. 3, 225–64

Equation (3.3) in fact states a little more. Although it is not possible to normalize null vectors in the usual way, the vectors u_μ have the property that they are covariantly constant along each of the generators. This fixes the *extents* of all the u_μ's in terms of the extent of any *one* of these u_μ's corresponding to each generator. On the other hand, there is no invariant way, which holds for all types of null hypersurface, of comparing the extents of the u_μ's on *different* generators.

Similar remarks apply to the spinor formalism. At each point of $\mathcal{N}$ we may choose a spinor ξ_A associated with u_μ as in (1.4), that is, $u_{AB'} = \xi_A \bar{\xi}_{B'}$. The phase of ξ_A is undetermined by this, but we may choose the flags so that they, also, are covariantly constant along each generator. That is,

$$u^\mu \partial_\mu \xi_A = \xi^B \bar{\xi}^{C'} \partial_{BC'} \xi_A = 0 \tag{3.4}$$

If this type of normalization for ξ_A is not chosen, then the right-hand side of (3.4) is not zero but a multiple of ξ_A. Contracting with ξ^A we get, in any case,

$$\xi^A \xi^B \bar{\xi}^{C'} \partial_{BC'} \xi_A = 0 \tag{3.5}$$

This is simply the general condition that the null vectors associated with the ξ_A's be tangent to geodesics. The stronger condition (3.4) on ξ_A will be adopted here, however, as it simplifies considerations of propagation along the generators. From (3.5) we get

$$\xi^A \bar{\xi}^{C'} \partial_{BC'} \xi_A = \rho \xi_B \tag{3.6}$$

$$\xi^A \xi^B \partial_{BC'} \xi_A = \sigma \bar{\xi}_{C'} \tag{3.7}$$

for some (possibly) complex numbers ρ, σ. In fact, it turns out that ρ is *real* and is a measure of the *convergence* of the generators at the point, while σ is complex and measures the *shear* of these generators [20, 21, 14]. Both quantities are pertinent to the intrinsic geometry of $\mathcal{N}$, as will be discussed shortly. The reality of ρ follows from (3.2):

$$0 = \bar{\xi}^{B'} \xi^C \{\partial_{AB'}(\xi_C \bar{\xi}_{D'}) - \partial_{CD'}(\xi_A \bar{\xi}_{B'})\} = (\rho - \bar{\rho}) \xi_A \bar{\xi}_{D'}$$

Also, it can easily be shown that assuming (3.4),

$$\bar{\xi}^{B'} \partial_{AB'} \xi^A = -\rho \tag{3.8}$$

whence

$$\rho = -\tfrac{1}{2} \partial_\mu u^\mu \tag{3.9}$$

Observe that if ξ_A is multiplied by a complex function λ of position in (3.6), (3.7), then ρ becomes $\rho \cdot \lambda\bar{\lambda}$ and σ becomes $\sigma \cdot \lambda^3 \bar{\lambda}^{-1}$. In each case the term involving derivatives of λ vanishes. Thus, ρ and σ possess a certain covariance with respect to such "gauge" transformations. Denoting the *weight* of ξ_A by (1, 0) and that of $\bar{\xi}_A$ by (0, 1), then ρ has weight (1, 1) and σ has weight

(3, −1). In general, if a quantity has weight (a, b), then $\frac{1}{2}(a+b)$ tells us how the quantity is affected by a change in extent of the null vector u_μ, while $\frac{1}{2}(a-b)$ tells how it depends on the direction of the flag determined by ξ_A. Both ρ and σ are therefore scaled in proportion to the extent of u_μ; ρ is independent of the direction of the flag but σ is dependent on it, any two opposite positions of the flag giving the same value for σ. Note that the null datum (2.12) for a field $\phi_A^{P'\cdots T'}{}_{D}$ with p primed and q unprimed indices is a quantity of weight (q, p). We shall see later that $\frac{1}{2}|q-p|$ is the spin of a corresponding "massless free field." This suggests that σ may be related to a spin-2 quantity, a fact that will become apparent later.

Let us now consider the geometrical meaning of ρ and σ in relation to the metric in $\mathfrak{N}$. At each point P of $\mathfrak{N}$ choose η_A with

$$\xi_A \eta^A = 1 \tag{3.10}$$

Then the vectors $u_\mu, v_\mu, \bar{v}_\mu$ with

$$u_{AB'} = \xi_A \bar{\xi}_{B'}, \qquad v_{AB'} = \xi_A \bar{\eta}_{B'}, \qquad \bar{v}_{AB'} = \eta_A \bar{\xi}_{B'}$$

span the tangent space to $\mathfrak{N}$ at P. The position vector relative to P of any neighboring point Q on $\mathfrak{N}$ is

$$\mathbf{PQ} \approx \bar{z} v_\mu + z \bar{v}_\mu + h u_\mu \tag{3.11}$$

to first order, z and h $(=\bar{h})$ being small. The spacelike interval PQ is, to first order,

$$PQ \approx 2^{1/2} |z| \tag{3.12}$$

The geometry of the neighborhood of P in $\mathfrak{N}$ is the product of the real Euclidean plane spanned by $v_\mu, \bar{v}_\mu$, and the null line given by u_μ. The u_μ component of the separation between two neighboring points does not contribute to their distance apart. The plane spanned by $v_\mu, \bar{v}_\mu$ acts as an Argand plane for the complex number $2^{1/2} z$.

To higher order, we must examine how these neighborhoods are related to each other. Let p, q be the generators through P, Q, respectively. Choose η_A, and therefore v_μ, v_μ, to be covariantly constant along p. We shall investigate how z changes [16] as P, Q vary along p, q.

Define

$$D \equiv u^\mu \partial_\mu \equiv \xi^A \bar{\xi}^{B'} \partial_{AB'} \tag{3.13}$$

and

$$\delta \equiv v^\mu \partial_\mu \equiv \xi^A \bar{\eta}^{B'} \partial_{AB'}, \qquad \bar{\delta} \equiv v^\mu \partial_\mu \equiv \eta^A \bar{\xi}^{B'} \partial_{AB'} \tag{3.14}$$

[16] Essentially this analysis was first carried out by Sachs [21].

 Null hypersurface initial data for classical fields of arbitrary spin and for general relativity. Widely circulated preprint, *Cornell* (1961). Also circulated in P. G. Bergmann's ARL Technical Documentary Report (1963) 63–56; *Gen. Relativity Gravitation* **12** (1980), no. 3, 225–64

Then

$$D\xi^A = 0, \qquad D\eta^A = 0 \tag{3.15}$$

and

$$\xi^A \delta\xi_A = \sigma, \qquad \xi^A \bar{\delta}\xi_A = \rho \tag{3.16}$$

by (3.7), (3.6), (3.10). Let y be a real function on $\mathfrak{N}$ which is constant along each generator:

$$Dy = 0 \tag{3.17}$$

The value of y at Q differs from that at P by an amount

$$(\bar{z}\delta + z\bar{\delta} + hD)y = \bar{z}\delta y + z\bar{\delta}y$$

to first order, by (3.11), (3.13), (3.14), (3.17), whence

$$D(\bar{z}\delta y + z\bar{\delta}y) = 0 \tag{3.18}$$

But from (3.15)

$$\begin{aligned} D\delta y &= D(\xi^A \bar{\eta}^{B'} \partial_{AB'} y) = \xi^A \bar{\eta}^{B'} (\xi^C \bar{\xi}^{D'} \partial_{CD'}) \partial_{AB'} y \\ &= \xi^A \bar{\eta}^{B'} \xi^C \bar{\xi}^{D'} \partial_{AB'} \partial_{CD'} y \\ &= \delta D y - (\delta \xi^C) \bar{\xi}^{D'} \partial_{CD'} y - \xi^C (\delta \bar{\xi}^{D'}) \partial_{CD'} y \end{aligned} \tag{3.19}$$

since covariant derivatives commute when acting on a scalar. From (3.17), (3.13) and (3.10), (3.14), we get

$$\xi^C \partial_{CD'} y = \bar{\xi}_{D'} \delta y, \qquad \bar{\xi}^{D'} \partial_{CD'} y = \xi_C \bar{\delta} y$$

so that (3.19) becomes

$$D\delta y = \sigma\bar{\delta}y + \rho\delta y$$

by (3.17), (3.16). Thus (3.18) gives

$$Dz = -\rho z - \sigma\bar{z} \tag{3.20}$$

since δy and $\bar{\delta} y$ are independent.

Equation (3.20) exhibits the fact that ρ measures the convergence of the generators and σ measures their shear. This is perhaps made a little clearer by the consequent formula

$$D(2^{1/2}|z|) = 2^{1/2}|z| \left\{ -\rho - \frac{1}{2}\left(\sigma \frac{\bar{z}}{z} + \bar{\sigma}\frac{z}{\bar{z}}\right)\right\}$$

since $2^{1/2}|z|$ is the distance PQ [see (3.12)]. The maximum convergence of p, q occurs when $\sigma\bar{z}/z$ is real and positive; the minimum (or maximum divergence) when $\sigma\bar{z}/z$ is negative. These values of z correspond to two directions at right

angles. The average convergence is measured by ρ. The presence of shear means that a small spacelike circle on $\mathfrak{N}$ becomes elliptical as it follows the generators.

The area A of a small triangle PQ_1Q_2 is given by

$$A = \frac{i}{2}(z_1\bar{z}_2 - z_2\bar{z}_1)$$

where z_i corresponds to Q_i as in (3.11). Allowing P, Q_1, Q_2 to vary along generators on $\mathfrak{N}$, we get, by (3.20)

$$DA = -2\rho A \tag{3.21}$$

Formula (3.21) clearly generalizes to any element of area cut off by a small tube of generators on $\mathfrak{N}$. The quantity -2ρ measures the rate of spreading of this area along the tube [compare (3.9)].

In flat space, since p and q are then straight lines, the vector **PQ** varies linearly with P and Q. Thus,

$$D^2 z = 0 \tag{3.22}$$

[this is generalized for curved space in (3.26)] and using (3.20) we get

$$z = (1 - l\rho_0)z_0 - l\sigma_0 z_0$$

The subscript 0 denotes the value of $l = 0$, where l is a parameter on each generator such that

$$D \equiv \frac{\partial}{\partial l} \tag{3.23}$$

(when the operators act on scalars). This means that for some point O on p, the spinor equivalent $p^{AB'}$ of the vector OP is

$$p^{AB'} = l\xi^A \bar{\xi}^{B'} \tag{3.24}$$

for each P on p. From (3.20) we also get

$$\sigma = \frac{\sigma_0}{(1 - l\rho_0)^2 - l^2\sigma_0\bar{\sigma}_0} \qquad \rho = \frac{\rho_0 + l(\sigma_0\bar{\sigma}_0 - \rho_0^2)}{(1 - l\rho_0)^2 - l^2\sigma_0\bar{\sigma}_0} \tag{3.25}$$

A null cone in flat space is characterized by the fact that $\sigma = 0$ and $\rho \neq 0$ everywhere on $\mathfrak{N}$. The vertex then occurs at $l = \rho_0^{-1}$, at which point ρ becomes infinite. A null hyperplane is given when $\rho = 0 = \sigma$ and this is the only case where $\mathfrak{N}$, if complete, is free from singularities. In the general case in flat space, most generators of $\mathfrak{N}$ encounter two places where ρ and σ become infinite:

$$l = (\rho_0 \pm |\sigma_0|)^{-1}$$

and there will also be other singularities where $\mathfrak{N}$ crosses itself.

 Null hypersurface initial data for classical fields of arbitrary spin and for general relativity. Widely circulated preprint, *Cornell* (1961). Also circulated in P. G. Bergmann's ARL Technical Documentary Report (1963) 63–56; *Gen. Relativity Gravitation* **12** (1980), no. 3, 225–64

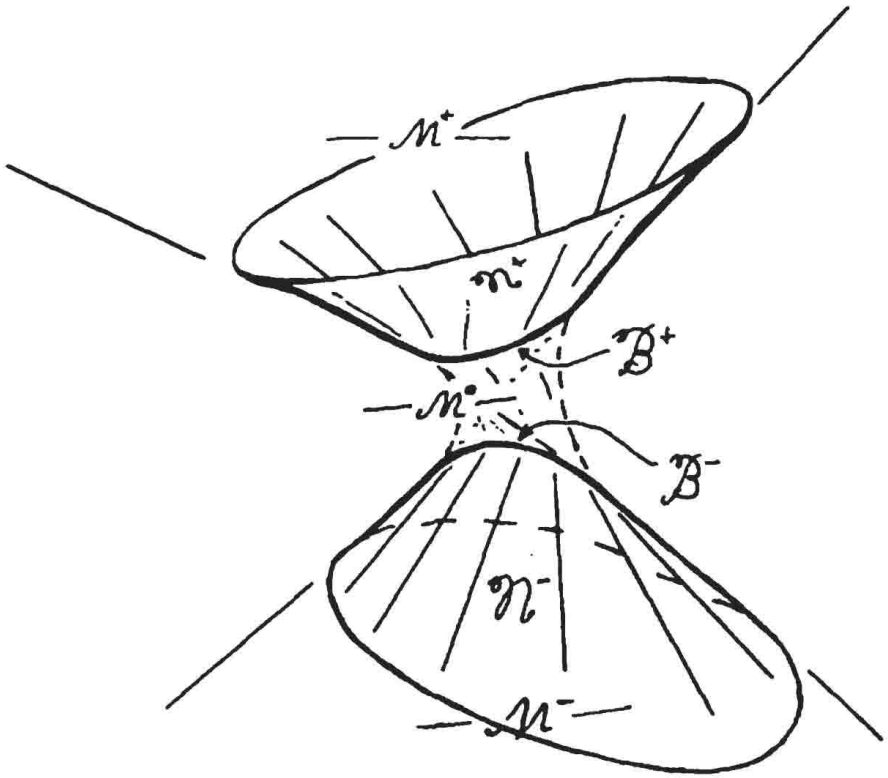


Fig. 2. A null hypersurface in a (2 + 1)-dimensional flat space-time. In (3 + 1) dimensions, $\mathscr{N}^+$ and $\mathscr{N}^-$ would be three dimensional and $\mathscr{B}^+$, $\mathscr{B}^-$ would in general be two dimensional (and branching). $\mathscr{N}^+$ and $\mathscr{N}^-$ divide the space-time into three regions: $\mathscr{M}^+$, $\mathscr{M}^0$, and $\mathscr{M}^-$.

The type of null hypersurface that will mainly be of interest here is one that is nonsingular in the remote past and future (see Figure 1). Then (except in the case of null hyperplanes, which are to be excluded here) the remote past region can be extended forwards in time until singularities are encountered, giving a maximal nonsingular region $\mathscr{N}^-$. Its boundary $\mathscr{B}^-$, consisting of singular points of $\mathscr{N}$, will be generally two-dimensional, branching, and spacelike. Similarly the remote future region can be extended backwards to obtain a maximal nonsingular region $\mathscr{N}^+$ with boundary $\mathscr{B}^+$. There will in general be other regions of $\mathscr{N}$ between $\mathscr{B}^-$ and $\mathscr{B}^+$. In the case of a null cone, however, $\mathscr{B}^-$ and $\mathscr{B}^+$ coincide at the vertex of $\mathscr{N}$. In the general case, $\mathscr{N}^+$ and $\mathscr{N}^-$ have the characteristic property that all their spacelike cross sections are *convex* surfaces.

In the next section, the determination of certain fields in special relativity from their null data by means of explicit integrals will be considered. For this purpose it will be supposed that the relevant null hypersurface is as just described above and that the null data are specified on $\mathscr{N}^+$ and on $\mathscr{N}^-$ only. It will also be necessary to specify the fields themselves on $\mathscr{B}^+$ and $\mathscr{B}^-$. In the case of a null cone, however, this last question does not arise because the fields at the vertex are determined by the null data at the vertex.

In curved space, in place of (3.22) we have the geodesic deviation equation:

$$D^2 z = -\omega z - \psi \bar{z} \tag{3.26}$$

where

$$\psi = R_{\alpha\beta\mu\nu} u^\alpha v^\beta u^\mu v^\nu$$

and

$$\omega = R_{\alpha\beta\mu\nu} u^\alpha v^\beta u^\mu \bar{v}^\nu = -\tfrac{1}{2} R_{\mu\nu} u^\mu u^\nu$$

the tensor $R_{\mu\nu\rho\sigma}$ being the Riemann curvature tensor and $R_{\mu\nu} = R^\alpha{}_{\mu\alpha\nu}$ the Ricci tensor [22, 21, 23]. The quantity ψ is complex and measures the purely astigmatic part of the geodesic deviation, while ω is real and measures the anastigmatic part.

In spinor form we have

$$\psi = \psi_{ABCD} \xi^A \xi^B \xi^C \xi^D \tag{3.27}$$

and

$$\omega = \phi_{ABC'D'} \xi^A \xi^B \bar{\xi}^{C'} \bar{\xi}^{D'}$$

where

$$\begin{aligned} R_{AE'BF'CG'DH'} = {} & \psi_{ABCD} \epsilon_{E'F'} \epsilon_{G'H'} + \epsilon_{AB} \epsilon_{CD} \bar{\psi}_{E'F'G'H'} \\ & + \tfrac{1}{12} R \{ \epsilon_{AC} \epsilon_{BD} \epsilon_{E'F'} \epsilon_{G'H'} + \epsilon_{AB} \epsilon_{CD} \epsilon_{E'H'} \epsilon_{F'G'} \} \\ & + \epsilon_{AB} \phi_{CDE'F'} \epsilon_{G'H'} + \epsilon_{CD} \phi_{ABG'H'} \epsilon_{E'F'} \end{aligned} \tag{3.28}$$

gives the decomposition of a general Riemann tensor into its irreducible parts. This generalizes (1.9). The spinor ψ_{ABCD} is totally symmetric and corresponds to the Weyl tensor $C_{\mu\nu\rho\sigma}$ as was mentioned in Section 1. The spinor $\phi_{ABC'D'}$ is symmetric in AB and in $C'D'$. It corresponds to the trace-free part of the Ricci tensor. The matter fields determine $\phi_{ABC'D'}$ and R (= $R^\alpha{}_\alpha$, the curvature scalar) directly through their energy tensor.

The spinor ψ_{ABCD} is considered here to represent the gravitational field and its null datum is ψ given by (3.27). The null data for the matter fields apparently determine the real number ω directly in many cases. For example, for the Einstein–Maxwell equations {see (2.7), (2.8), [13]} we have $\omega = k\phi\bar{\phi}$ with $\phi = \phi_{AB} \xi^A \xi^B$. Of course in empty space $\omega = 0$. The presence of a λ term in Einstein's equations does not affect any of this since λ contributes only to R.

Differentiating (3.20) and using (3.26), (3.20), we obtain

$$D\rho = \rho^2 + \sigma\bar{\sigma} + \omega \tag{3.29}$$

$$D\sigma = 2\rho\sigma + \psi \tag{3.30}$$

so that ω and ψ govern the rates of change of ρ and σ. From (3.30) it is clear that in general relativity a null cone cannot be expected, in general, to have

$\sigma = 0$ everywhere since the presence of ψ will destroy this. In fact, apart from at the vertex, a null cone possesses no *local* features that distinguish it from any other type of null hypersurface. This is one important reason why the study of the initial value problem on null hypersurfaces other than just null cones is of interest.

However, the presence of the vertex in the case of a null cone makes many considerations easier for this case. The vertex is a convenient reference point to which all spinor (or tensor) quantities may be referred by parallel transport along a geodesic, as was done in the last section. The gravitational null datum ψ (and similarly, also ω) may be given as a complex (and real) function of three real parameters, two of these parameters determining the direction of the relevant generator at the vertex and the third determining the extent of this generator. Corresponding to each null direction at O, one ξ^A spinor must be chosen. Then ξ^A at every other point of $\mathfrak{N}$ is determined by parallel transport. The interpretation of the complex number ψ (and also of ω) is then well defined at each point of $\mathfrak{N}$. (Of course ψ, ω, and also any other null data specified on $\mathfrak{N}$ will not generally be single valued at O. They will be functions of the two variables which determine the null directions at O.) If the two quantities ψ and ω are known everywhere on $\mathfrak{N}$, then equation (3.26) may be integrated (constants of integration being determined at the vertex) and so the intrinsic metric of $\mathfrak{N}$ is known, by (3.12). The quantities σ and ρ may also be obtained, from (3.29), (3.30) or from (3.20).

Conversely, knowledge of the intrinsic geometry of $\mathfrak{N}$ might be expected to be sufficient for the determination of ψ and ω. However, the concept of intrinsic geometry for a null hypersurface is not quite a straightforward matter, and, in fact, at least three different types of intrinsic geometry can be distinguished. This is because the metric tensor of $\mathfrak{N}$ has zero determinant and therefore no inverse. Christoffel symbols cannot then be defined which lead to a general notion of parallel transport. These matters will be discussed in detail elsewhere [24], but it turns out that if the intrinsic geometry of $\mathfrak{N}$ is considered to be determined by the metric of *alone*, then neither ψ nor ω are covariant quantities. That is to say, ψ and ω cannot in general be determined from the metric of alone. However, it turns out that the expression

$$\sigma\omega - \rho\psi$$

is a suitably covariant quantity, so that in *empty* space ($\omega = 0$), ψ *can* be determined from the metric of $\mathfrak{N}$ except in regions where $\rho = 0$. On the other hand, if the concept of intrinsic geometry for $\mathfrak{N}$ is considered to carry with it the concept of affine parameters on each generator of $\mathfrak{N}$, then in this case it turns out that *both* ψ and ω are covariant quantities and are thus determined by the intrinsic geometry of $\mathfrak{N}$ in this sense.

If we bear this in mind, one of the results of the last section can be interpreted as follows: given an analytic empty space-time $\mathfrak{M}$, the geometry of $\mathfrak{M}$

is uniquely determined[17] (except perhaps globally) by the intrinsic geometry (with the above provisos) of the null cone $\mathcal{N}$ of any one point[18] O in $\mathcal{M}$. The metric in $\mathcal{M}$ can be used as essentially independent (analytic) initial data for $\mathcal{M}$, although the use of ψ is probably more satisfactory. A third alternative would be to use σ, which again can be chosen arbitrarily. (See also [3].)

The analyticity condition would seem to be inessential, but this will require proof. Also, it will be of interest to generalize this result to null hypersurfaces $\mathcal{N}$ other than cones. For this case let there be an initial two-dimensional spacelike surface $\mathcal{B}^+$, say, which (locally) lies at the intersection of two pieces $\mathcal{N}_1^+, \mathcal{N}_2^+$ of $\mathcal{N}$ lying to the future of $\mathcal{B}^+$ and generated by the null geodesics which meet $\mathcal{B}^+$ and are orthogonal to it (see Figure 3). Initial data appropriate for the determination of the space-time $\mathcal{M}^+$ which lies between (i.e., to the future of) $\mathcal{N}_1^+$ and $\mathcal{N}_2^+$ can be described roughly as follows. Firstly, the intrinsic geometry of the Riemannian 2-space $\mathcal{B}^+$ must be given. This determines the Gaussian curvature of each point of $\mathcal{B}^+$.

Secondly, a certain imbedding invariant of $\mathcal{B}^+$ in $\mathcal{N}_1^+$ must be given at each point of $\mathcal{B}^+$. This quantity is a kind of extrinsic curvature for $\mathcal{B}^+$ and it combines naturally in the form of a complex number κ with the Gaussian curvature of $\mathcal{B}^+$. The definition and meaning of κ will be discussed elsewhere [24]. It turns out that it is symmetrical with respect to $\mathcal{N}_1^+$ and $\mathcal{N}_2^+$ and it determines how spinors $\xi^A_{(1)}, \xi^A_{(2)}$ for the various generators of $\mathcal{N}_1^+, \mathcal{N}_2^+$ are to be related to one another. Next, the initial $\rho_1, \sigma_1, \rho_2, \sigma_2$ (at $\mathcal{B}^+$) for $\mathcal{N}_1^+, \mathcal{N}_2^+$ are to be given.[19] Finally, the null datum ψ is to be given on $\mathcal{N}_1^+$ and on $\mathcal{N}_2^+$. It seems that all these quantities can be specified arbitrarily and together they determine $\mathcal{M}^+$ uniquely.

§(4): *The Free Field Integrals*

According to the point of view of this paper, a "free field" is a field $\theta^{P'\cdots T'}_{A\cdots D}$ in flat space, all of whose derivatives are symmetric in all primed and all unprimed indices:

$$\partial_E{}^{U'} \cdots \partial_G{}^{W'} \theta^{P'\cdots T'}_{A\cdots D} = \partial_{(E}{}^{(U'} \cdots \partial_G{}^{W'} \theta^{P'\cdots T')}_{A\cdots D)} \tag{4.1}$$

[17] It might be thought that since the arguments of Section 2 only applied to the Riemann tensor, the *metric* of $\mathcal{M}$ would not be necessarily determined by $\mathcal{N}$. However, since the metric and Christoffel symbols are known at O, and the curvature is known everywhere, there is no freedom left to $\mathcal{M}$ [25].

[18] It has been pointed out to me by A. Komar that if it is desired to characterize empty space-times by means of a nonredundant set of variables, this can apparently usually be done (locally) by "fixing" the point O and the vectors given by (1.7) in terms of invariants and principal directions in the space. Then the components of ψ_{ABCD} are determined at O and the remaining ψ variables give the nonredundant set.

[19] This ties in with the fact that in the flat-space cases it is necessary to specify the *field* on $\mathcal{B}^+$. For the required components of ψ_{ABCD} on $\mathcal{B}^+$ can, in fact, be simply expressed in terms of $\rho_1, \sigma_1, \rho_2, \sigma_2$ and their tangential derivatives [24].

 Null hypersurface initial data for classical fields of arbitrary spin and for general relativity. Widely circulated preprint, *Cornell* (1961). Also circulated in P. G. Bergmann's ARL Technical Documentary Report (1963) 63–56; *Gen. Relativity Gravitation* **12** (1980), no. 3, 225–64

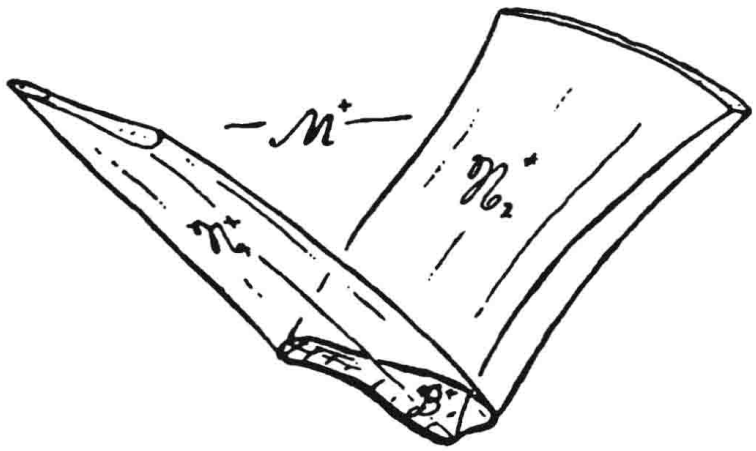


Fig. 3. A general characteristic initial value problem for general relativity. $\mathcal{N}_1^+$ and $\mathcal{N}_2^+$ are two null hypersurfaces which meet at a spacelike 2-surface $\mathcal{B}^+$. The space-time lying between them is to be determined from data on $\mathcal{N}_1^+$, $\mathcal{N}_2^+$, and $\mathcal{B}^+$.

To avoid confusion with the conventional nomenclature, however, these fields will be called *massless free fields.* Interactions of fields here concern skew parts (or contractions) of derivatives of fields. There are two forms of interaction to be considered, namely, those given by field equations and those given by commutator equations (see Section 2.) For a massless free field, neither type of interaction may occur, so derivatives, must commute and the space-time is flat. Also, according to the definition given here, all these free fields must, in fact, have zero rest mass. Fields of finite rest mass are considered as sets of interacting fields. Clearly a massless free field gives, by itself, an exact set.

Examples of massless free fields are the source-free Maxwell field ϕ_{AB} (as was seen in Section 2), the Weyl neutrino field ν_A and the linearized Einstein gravitational field ψ_{ABCD}. This last spinor is obtained from the linearized empty-space Riemann tensor by (1.9) [or (3.28)]. The field equations are all special cases of (1.11), as was remarked earlier, but now necessarily in flat space. These fields and their complex conjugates:

$$\partial^{AP'}\theta_{AB\cdots E} = 0 \text{ or } \partial_{AP'}\phi^{P'Q'\cdots T'} = 0 \tag{4.2}$$

for nonzero spin, together with the flat-space D'Alembert field:

$$\Box\chi = 0 \tag{4.3}$$

for zero spin, will be called the *basic* free fields. It is easy to verify that all the basic fields are in fact massless free fields according to the definition (4.1). This was done for the Maxwell field in (2.3) and the other cases are similar. [Equation (1.12) may be used for the D'Alembert case.]

The remaining massless free fields possess both primed and unprimed indices and the field equations are

$$\partial^{AU'}\theta^{P'\cdots T'}_{AB\cdots D} = 0, \qquad \partial_{EP'}\theta^{P'Q'\cdots T'}_{AB\cdots D} = 0 \tag{4.4}$$

However, these fields turn out to be simply derivatives of basic free fields (see also [26]), for from (4.1) we get

$$\partial_E{}^{U'}\theta_{A\cdots CD}^{P'\cdots S'T'} - \partial_D{}^{T'}\theta_{A\cdots CE}^{P'\cdots S'U'} = 0$$

This is just the spinor translation of a vanishing curl, whence

$$\theta_{A\cdots CD}^{P'\cdots S'T'} = \partial_D{}^{T'}\phi_{A\cdots C}^{P'\cdots S'} \tag{4.5}$$

Since $\theta_{A\cdots CD}^{P'\cdots S'T'}$ is symmetric in $P'\cdots S'$ and in $A\cdots C$, this symmetry may also be imposed on $\phi_{A\cdots C}^{P'\cdots S'}$ without changing (4.5). Clearly $\phi_{A\cdots C}^{P'\cdots S'}$ is also a massless free field according to (4.1). Repeating the argument until one or the other set of indices is exhausted, we finally arrive at a basic free field from which $\theta_{A\cdots D}^{P'\cdots T'}$ is obtained by repeated uncontracted differentiation.

The remainder of this section will be concerned with an explicit formula which expresses any basic free field $\theta_{AB\cdots E}$ with n indices at an arbitrary point P of space-time, in terms of the appropriate initial data on a given null hypersurface $\mathfrak{N}$. Throughout this section it will be assumed that the space-time is flat and that $\mathfrak{N}$ is as described in Section 3 (see Figure 2), having a maximal nonsingular future part $\mathfrak{N}^+$ and nonsingular past part $\mathfrak{N}^-$, bounded respectively by sets $\mathfrak{B}^+$ and $\mathfrak{B}^-$ of points singular for $\mathfrak{N}$. For convenience, the union $\mathfrak{B}^+ \cup \mathfrak{B}^-$ will be denoted by $\mathfrak{B}$. The null datum θ is to be specified on $\mathfrak{N}^+$ and $\mathfrak{N}^-$ only, and the whole field $\theta_{AB\cdots E}$ is to be specified on $\mathfrak{B}$ (that is, both on $\mathfrak{B}^+$ and on $\mathfrak{B}^-$). The space-time to the future of $\mathfrak{N}^+$ is $\mathfrak{M}^+$, that to the past of $\mathfrak{N}^-$ is $\mathfrak{M}^-$ and the space-time between $\mathfrak{N}^+$ and $\mathfrak{N}^-$ is $\mathfrak{M}^0$.

Considering, for the moment, points P in $\mathfrak{M}^+$ or $\mathfrak{M}^-$ only, whose null cone does not meet $\mathfrak{B}$, the fundamental formula (to be proved shortly) is

$$\theta_{AB\cdots E}(P) = \frac{1}{2\pi}\int_{\mathcal{S}} \frac{1}{r}\zeta_A\zeta_B\cdots\zeta_E\{D\theta - (n+1)\theta\rho\}\,d\mathcal{S} \tag{4.6}$$

The integral is to be taken over all points Q that lie on the two-dimensional intersection $\mathcal{S}$ of the null cone $\mathcal{C}$ of P with $\mathfrak{N}^+$ or $\mathfrak{N}^-$ (see Figure 4). The set $\mathcal{S}$ is compact and connected and has the topology of a sphere. Every null line through P meets $\mathcal{S}$ in one and only one point.

On the other hand, if P were to lie in $\mathfrak{M}^0$ then $\mathcal{S}$ would extend to infinity and consist of at least two disjoint pieces, one or more on each of $\mathfrak{N}^+$, $\mathfrak{N}^-$. Thus, in this case, the application of (4.6) would involve some considerations of convergence. Moreover, although *almost* every null line through P would meet $\mathcal{S}$ in one and only one point, there would also be one-dimensional set of exceptional null lines through P *not* meeting $\mathcal{S}$ in any *finite* point. It will turn out [see (4.49)] that if P lies in $\mathfrak{M}^0$, there is, in general, an extra contribution to (4.6) from these null lines. However, if a certain condition of "continuity at infinity" is satisfied by θ (and $\mathcal{C}$ does not meet $\mathfrak{B}$), then (4.6) can still be expected to hold true [see (4.48)].

 Null hypersurface initial data for classical fields of arbitrary spin and for general relativity. Widely circulated preprint, *Cornell* (1961). Also circulated in P. G. Bergmann's ARL Technical Documentary Report (1963) 63–56; *Gen. Relativity Gravitation* **12** (1980), no. 3, 225–64

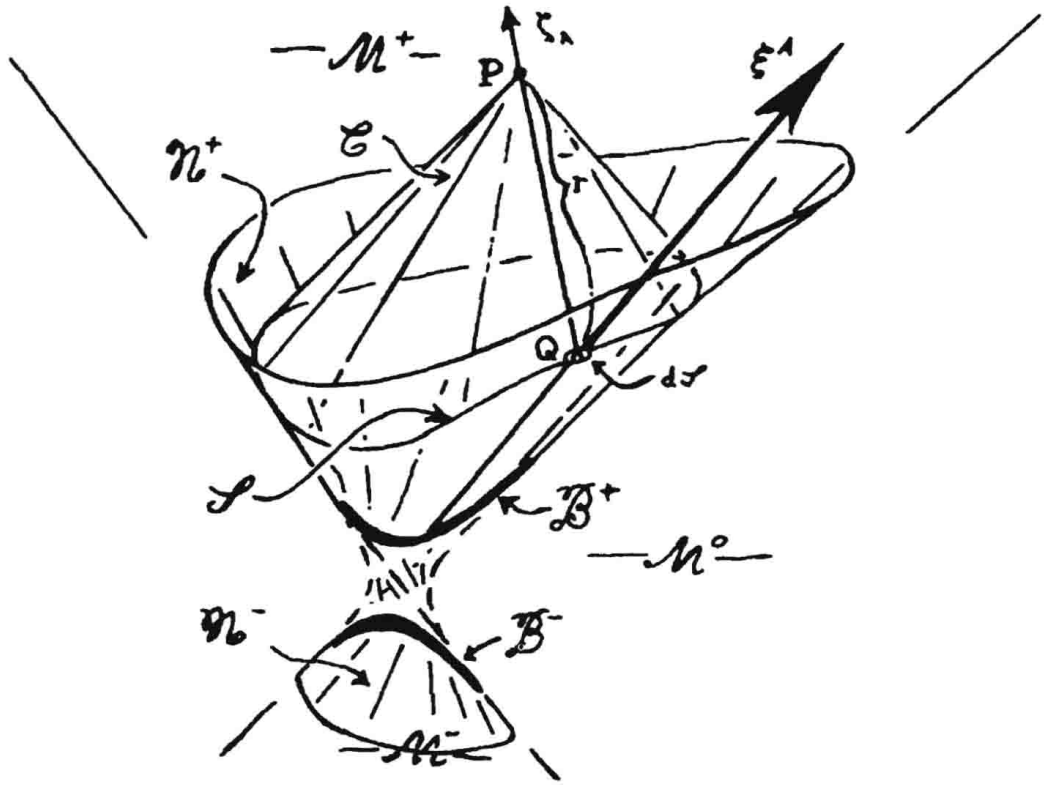


Fig. 4. The basic free field at the event P, lying in $\mathscr{M}^+$, can be obtained as an integral over the intersection $\mathcal{S}$ of the null cone of P with $\mathscr{I}^+$ (here, it does not meet $\mathscr{B}^+$). In (3 + 1) dimensions $\mathcal{S}$ would be two dimensional, with the topology of a sphere.

The meaning of θ in (4.6) is, of course [see (2.12)]

$$\theta = \theta_{AB\cdots E}\xi^A \xi^B \cdots \xi^E$$

where ξ^A is a spinor associated with $\mathscr{N}$ at Q in accordance with Section 3. Thus, $D\xi^A = 0$ where the operator

$$D \equiv \xi^F \bar{\xi}^{P'} \partial_{FP'}$$

is, as in (3.13), a differentiation in the null direction in $\mathscr{N}$ at Q. The integer n, as was mentioned earlier, is the number of indices of $\theta_{A\cdots E}$ (i.e. twice the spin). The quantity ρ measures the convergence of the generators of $\mathscr{N}$ at Q according to (3.6), (3.8), or (3.9). It is negative for $\mathscr{N}^+$ and positive for $\mathscr{N}^-$. The spinor ζ^A is normalized against ξ^A:

$$\zeta_A \xi^A = 1 \tag{4.7}$$

and corresponds to the null direction **QP**, the spinor equivalent of the vector **QP** being

$$r\zeta^A \bar{\zeta}^{B'}$$

Thus, r measures the extent of QP, scaled in terms of $\zeta^A \bar{\zeta}^{B'}$. It is positive (negative) if P lies to the future (past) of Q. The differential $d\delta$ is the element of surface area of δ at Q. (It will be observed that the integral (4.6) is invariant under any scale transformation $\xi^A \to \lambda\xi^A$, where $\lambda(\neq 0)$ is constant along each generator of $\mathfrak{N}$, since

$$\theta \to \lambda^n\theta, \quad D \to \lambda\bar{\lambda}D, \quad \rho \to \lambda\bar{\lambda}\rho, \quad \zeta^A \to \lambda^{-1}\zeta^A, \quad r \to \lambda\bar{\lambda}r, \quad d\delta \to d\delta$$

Now consider the cases when $\mathcal{C}$ *does* meet $\mathfrak{B}$. These cases can, in fact, be regarded as limiting cases of situations in which (4.6) applies, but care must be taken so as to obtain the correct interpretation in the limit. The cone $\mathcal{C}$ meets $\mathfrak{B}$ in a set $\mathfrak{E}$ of points which are singular for the surface δ. Since $\mathfrak{B}^+$ and $\mathfrak{B}^-$ are, in the main, two dimensional, $\mathfrak{E}$ consists of singular curves or "edges" on δ, together with (possibly) some "vertices" (generally occurring where these curves meet). Thus, δ becomes a kind of curvilinear polyhedron (see Figure 5).

Such a surface δ can always be approximated arbitrarily closely on $\mathcal{C}$ by a *smooth* surface δ' which is everywhere spacelike. The null lines which meet δ' orthogonally generate two null hypersurfaces, one of which is $\mathcal{C}$, the other, $\mathfrak{N}'$, being nonsingular in the neighborhood of δ'. The formula (4.6) can be applied to $\mathfrak{N}'$, in place of $\mathfrak{N}$, so allowing δ' to approach δ, we obtain the correct formula for case of $\mathfrak{N}$. It may be remarked, however, that the limiting position $\mathfrak{N}^*$ of $\mathfrak{N}'$ will not be identical with $\mathfrak{N}$ everywhere. The parts of $\mathfrak{N}$ generated by null lines that meet δ will in fact be parts of $\mathfrak{N}^*$. But $\mathfrak{N}^*$ will also contain connecting pieces of null cones whose vertices are the vertices of the curvilinear polyhedron δ, and connecting pieces of other null hypersurfaces that are singular along the singular curves of δ. These pieces of $\mathfrak{N}^*$ will all have infinite values of ρ at the singular places on δ, so (4.6) needs examining closely here.

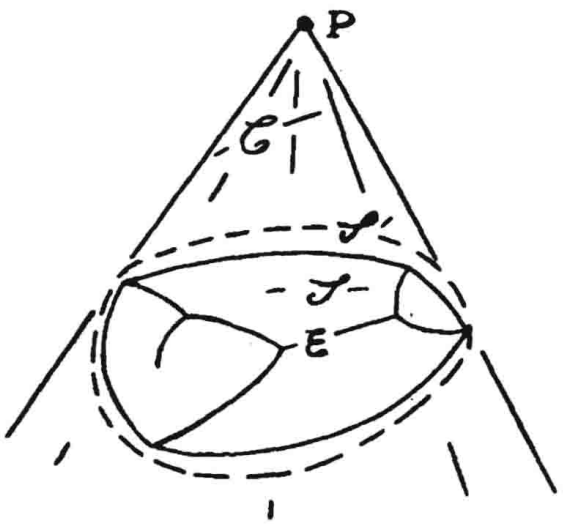


Fig. 5. When the null cone $\mathcal{C}$ of P meets $\mathfrak{B}$, δ becomes a kind of curvilinear polyhedron. The edges and vertices of δ are the places, $\mathfrak{E}$, where $\mathcal{C}$ meets $\mathfrak{B}$.

The integral (4.6) may be expressed as a sum of two parts, one involving $D\theta$ and not involving ρ and the other involving $\rho\theta$. Since the dimension of the singular region $\mathfrak{E}$ is less than that of $\mathcal{S}$, there will be no contribution to the first part of the integral at $\mathfrak{E}$, the integrand being finite. A way of examining the second part is to rewrite it as

$$\frac{n+1}{2\pi}\int_{\mathcal{K}} \zeta_A \zeta_B \cdots \zeta_E \theta \, d\mathcal{K} \tag{4.8}$$

where

$$d\mathcal{K} = -\rho \cdot \frac{1}{r} d\mathcal{S}$$

turns out to be the (signed) element of surface area (at a point corresponding to Q) of the surface $\mathcal{K}$ described by $\xi^A \zeta_B$, as Q varies, in the four-dimensional complex (vector) space $\mathfrak{Q}$ of spinors $q^A{}_B$. The metric in $\mathfrak{Q}$ is determined from a definition of scalar product for two elements $q_{(1)}{}^A{}_B, q_{(2)}{}^A{}_B$ as the complex number $q_{(1)}{}^A{}_B \, q_{(2)}{}^B{}_A$. It can be shown[20] that the quantity $-\rho/r$ is just half the Gaussian curvature of $\mathcal{S}$ at Q, so we expect finite contributions to (4.8) from places on $\mathcal{S}$ where this curvature involves delta functions. Such places will occur on the singular region $\mathfrak{E}$ of $\mathcal{S}$ (but unlike the case of an ordinary plane-faced polyhedron in Euclidean space, it is the edges and *not* the vertices[21] of $\mathcal{S}$ that give contributions). The integration of (4.8), of course, does not involve delta functions, but θ has to be suitably interpreted.

If the validity of (4.6) is now assumed for the cases when $\mathcal{S}$ is nonsingular, by carrying out the above limiting procedure in detail the following formula may be obtained. It will hold for all points P in $\mathfrak{M}^+$ or $\mathfrak{M}^-$ [and under a condition of "continuity at infinity" on θ, for all P in $\mathfrak{M}^0$ not on $\mathcal{N}$, see (4.48); otherwise (4.50) must be added]:

$$\theta_{AB\cdots E}(P) = \frac{1}{2\pi}\int_{\mathcal{S}-\mathfrak{E}} \frac{1}{r}\zeta_A \zeta_B \cdots \zeta_E \{D\theta - (n+1)\rho\theta\} \, d\mathcal{S}$$

$$+ \frac{(n+1)}{2\pi 2^{1/2}}\int \frac{1}{r}\zeta_A \zeta_B \cdots \zeta_E \theta^* \, d\mathfrak{E} \tag{4.9}$$

[20] See [24]. Such a definition of a surface $\mathcal{K}$ applies also to a *general* surface $\mathcal{S}$ in Minkowski space. Both $\mathcal{C}$ and $\mathcal{N}$ are then general null hypersurfaces and the relation $d\mathcal{K} = \kappa \, d\mathcal{S}$ holds where $\kappa = \sigma\sigma^* - \rho\rho^*$ is the *complex* curvature quantity for $\mathcal{S}$, that was mentioned at the end of Section 3, ρ^* and σ^* being the convergence and shear corresponding to $\mathcal{C}$. Thus, $\mathcal{K}$ generalizes, for an arbitrary surface in Minkowski space, Gauss' concept of the "sphere of unit normals" for a surface in Euclidean space. It is possible also to define $\mathcal{K}$ for surfaces in curved space–time.

[21] Any curvilinear polyhedron that lies in a nonsingular region of a null hypersurface (such as $\mathcal{C}$ here) has the property that the angles at any vertex must always sum to 2π. This follows at once from the nature of the neighborhood of a point in a null hypersurface. Thus, no finite contribution to the curvature of $\mathcal{S}$ can arise from the vertices.

with

$$\theta^* = \pm |\, \xi_{(0)}{}^A \xi_{(1)A} \,| \cdot \int_0^1 \theta_{AB\cdots E} \xi_{(t)}{}^A \xi_{(t)}{}^B \cdots \xi_{(t)}{}^E \, dt \qquad (4.10)$$

$$\xi_{(t)}{}^A = t\xi_{(1)}{}^A + (1-t)\xi_{(0)}{}^A \qquad (4.11)$$

The spinors $\xi_{(0)}{}^A, \xi_{(1)}{}^A$ are the ξ^A spinors corresponding to the two branches of $\mathfrak{N}$ which intersect in the relevant portion of $\mathfrak{B}$. They must both be normalized against ζ_A:

$$\zeta_A \xi_{(0)}{}^A = 1 = \zeta_A \xi_{(1)}{}^A$$

This can be done because of the freedom $\xi^A \to \lambda\xi^A$. (Hence ζ_A enters indirectly into the definition of θ^*.) The parts of $\mathfrak{E}$ (vertices) that lie on three or more branches (or on only one branch) of $\mathfrak{N}$ form a subset of $\mathfrak{E}$ of measure zero and can therefore be ignored. The differential $d\mathfrak{E}$ in (4.9) is the element of proper length of $\mathfrak{E}$ at Q. The ambiguous sign in (4.10) is to be taken as "+" if Q lies on $\mathfrak{B}^+$ and as "−" if Q lies on $\mathfrak{B}^-$.

Although formula (4.9) has been stated in terms of a particular kind of null hypersurface with a particular arrangement of initial data, it has, in fact, a somewhat wider domain of application. For example, suppose two null hyperplanes $\mathfrak{N}_1$ and $\mathfrak{N}_2$ intersect in a plane $\mathfrak{B}^+$ and suppose the null datum θ is given on the parts $\mathfrak{N}_1^+$ $\mathfrak{N}_2^+$ of $\mathfrak{N}_1, \mathfrak{N}_2$, respectively, which lie to the future of $\mathfrak{B}^+$, and that the field $\theta_{AB\cdots E}$ is given on $\mathfrak{B}^+$. Then (4.9) [with + sign in (4.10)] holds for points P that lie to the future of both $\mathfrak{N}_1^+$ and $\mathfrak{N}_2^+$. The formula (4.9) can also be applied (provided θ vanishes suitably fast at infinity) to points P that lie to the past of $\mathfrak{N}_1$ or $\mathfrak{N}_2$ (it being still true that *almost* every null line though P meets $\mathfrak{N}_1^+, \mathfrak{N}_2^+$, or $\mathfrak{B}^+$), but in this case the field is not uniquely determined at P by the given data so (4.9) gives only a particular field consistent with the data. This would also apply to the case of giving the null datum, on, say, the whole of $\mathfrak{N}_1$ only. The reason for the particular choice of arrangement of null hypersurface and data that is being made here is that the field is apparently always uniquely determined *throughout* space-time by the data and that the data can be given arbitrary values.[22] The essential feature of such an arrangement seems to be that every null line (excluding the generators of $\mathfrak{N}$) meets the region, on which the data are specified, in just one point, or else it "touches" it at infinity. (A null line l touches $\mathfrak{N}$ at infinity if the null hyperplane containing l also contains a

[22] There will be certain conditions at infinity to be satisfied by the null datum, however, to ensure convergence of the integrals for points P in $\mathfrak{N}^0$. It should be noted, in addition, that even if these conditions are met and if also the question of "continuity at infinity" has been dealt with [see (4.48), (4.50)], there still remains the fact that continuous, differentiable and reasonable-looking data can produce a *discontinuous* field, the discontinuity lying along $\mathfrak{N}$. This is because the limit of the region of integration as P approaches $\mathfrak{N}$ from either side, is quite different in the two cases.

generator of $\mathcal{N}$.) The geometrical-topological proof that the regions considered here (see Figure 2) in fact have this property, is essentially simple but will not be given here.

Let us now proceed to the proof of (4.6) [or (4.9)] where for simplicity it will be assumed at first that the given data are zero in the neighborhood of $\mathcal{B}$ so that (4.6) will apply to *all* points P in $\mathcal{M}^+$ or $\mathcal{M}^-$. The proof will be in three parts. Firstly it will be shown that the field $\theta_{AB\dots E}(P)$ defined from θ by the integral (4.6) satisfies the equation for a basic free field. Secondly it will be shown that this field agrees with θ on $\mathcal{N}^+$ and $\mathcal{N}^-$, and thirdly a uniqueness theorem for massless free fields with the given initial data will be proved.

For the first part of the proof, the integral (4.6) must be differentiated. Let $p^{AB'}$ be the *position spinor* of the point P (that is, $p^\mu = \sigma^\mu{}_{AB'}p^{AB'}$ is the position vector of P). Then Q is a function of $p^{AB'}$ and of the generator g of $\mathcal{N}$ on which it lies. The spinor ξ^C is a function of g but, being constant along g, is independent of $p^{AB'}$. On the other hand, ζ_C, r, ρ, θ, and $D\theta$ are all functions of both g and $p^{AB'}$. Define

$$\Delta_{AB'} \equiv \frac{\partial}{\partial p^{AB'}}$$

Then

$$\Delta_{AB'}\xi^C = 0$$

and since $\xi_C\zeta^C = -1$,

$$\xi_C \Delta_{AB'}\zeta^C = 0 \tag{4.12}$$

Let l be a parameter on each generator g of $\mathcal{N}$ such that $D \equiv \partial/\partial l$ [see (3.23)]. Let the position spinor of the point on g for which $l = 0$ be $h^{AB'}$. Then (see Figure 6) we have

$$p^{AB'} = h^{AB'} + l\xi^A\bar{\xi}^{B'} + r\zeta^A\bar{\zeta}^{B'} \tag{4.13}$$

Differentiating with respect to $p^{FR'}$ we get

$$\delta^A_F\delta^{B'}_{R'} = \xi^A\bar{\xi}^{B'}\Delta_{FR'}l + \zeta^A\bar{\zeta}^{B'}\Delta_{FR'}r + r\zeta^A\Delta_{FR'}\bar{\zeta}^{B'} + r\bar{\zeta}^{B'}\Delta_{FR'}\zeta^A \tag{4.14}$$

Contracting this with $\zeta_A\bar{\zeta}_{B'}$, we obtain [using (4.7)]

$$\Delta_{FR'}l = \zeta_F\bar{\zeta}_{R'} \tag{4.15}$$

Contracting (4.14) with $\xi_A\bar{\xi}_{B'}$ and using (4.12) we get

$$\Delta_{FR'}r = \xi_F\bar{\xi}_{R'} \tag{4.16}$$

Also, since (4.12) tells us $\Delta_{AB'}\zeta^C$ is of the form $a_{AB'}\xi^C$, contracting (4.14) with $\zeta_A\bar{\xi}_{B'}$ gives us

$$\Delta_{FR'}\zeta^A = -\frac{1}{r}\zeta_F\bar{\xi}_{R'}\xi^A \tag{4.17}$$

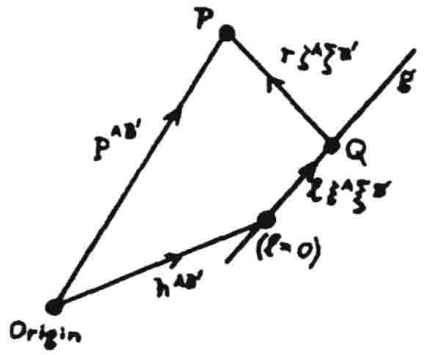


Fig. 6. Space-time diagram for the relation (4.13).

From (4.15), if ϕ is any quantity defined at Q, we have

$$\Delta_{FR'}\phi = \zeta_F\bar{\zeta}_{R'}D\phi$$

Hence, by (3.29),

$$\Delta_{FR'}\{D\theta - (n+1)\rho\theta\} = \zeta_F\bar{\zeta}_{R'}\{D^2\theta - (n+1)\rho D\theta - (n+1)(\rho^2 + \sigma\bar{\sigma})\theta\} \tag{4.18}$$

Also, by (3.21) (*et seq.*),

$$\Delta_{FR'}\,d\delta = -\zeta_F\bar{\zeta}_{R'}\cdot 2\rho\,d\delta \tag{4.19}$$

Thus, applying the operator $\Delta_F{}^{R'}$ to the expression (4.6) and using (4.16), (4.17), (4.18), (4.19) and the fact that the operators $\Delta_{FR'}$ and $\partial_{FR'}$ are the same when applied to a qunatity defined at P, we obtain

$$\partial_F{}^{R'}\theta_{AB\cdots E}(P) = \frac{1}{2\pi}\int_{\delta}\left[\frac{1}{r}\zeta_A\zeta_B\cdots\zeta_E\zeta_F\bar{\zeta}^{R'}\{D^2\theta - (n+3)\rho D\theta + (n+1)\right.$$

$$\left.\cdot(\rho^2 - \sigma\bar{\sigma})\theta\} - \frac{(n+1)}{r^2}\zeta_{(A}\zeta_B\cdots\zeta_E\xi_{F)}\bar{\xi}^{R'}\{D\theta - (n+1)\rho\theta\}\right]d\delta \tag{4.20}$$

This expression is manifestly symmetric in $AB\cdots EF$, so that, provided $n > 0$, the appropriate field equation (4.2) is satisfied as required. For the case $n = 0$ (zero spin), (4.20) must be differentiated once further [see (4.21) with $n = 0$, $k = 2$]. This again is suitably symmetric so the required field equation (4.3) is satisfied.

It will be seen from the above argument that the *particular* form of the factor $\{D\theta - (n+1)\rho\theta\}\,d\delta$ given in (4.6) is not essential to this stage of the proof.[23] It enters only at the second stage when the calculated field is compared with the null datum on $\mathfrak{N}$. In fact, (4.18), (4.19) and thus (4.20), were expressed in somewhat more detail than was absolutely necessary. However, formula (4.20) is of interest in itself since it indicates the analog of (4.6) for the

[23] In fact, this argument amounts essentially to a proof of the fact that $\zeta_A\zeta_B\cdots\zeta_E\Delta$ is a solution of the basic free-field equation, where Δ is the usual Dirac singular solution of $\Box\Delta = 0$.

massless free field $\phi_A^{R'}{}_{\dots EF} = \partial_F{}^{R'}\theta_{A\dots E}$ with null datum $\phi = D\theta$. It tells us, furthermore, that [assuming (4.6)] the field $\phi_A^{R'}{}_{\dots F}$ at the point P depends on its null datum ϕ not only on the 2-surface $\mathcal{S}$, but also on regions of $\mathcal{N}$ that lie in the *interior* of the null cone of P. This is because θ appears undifferentiated in two places in the integrand in (4.20), and to calculate a suitable θ from ϕ an integration along the generators of $\mathcal{N}$ is required. This process is simple enough when $\mathcal{N}$ is a null cone since the constants of integration are all determined by the (arbitrary) value of ϕ at the vertex. It is a little more involved when $\mathcal{N}$ is a general null hypersurface. In any case, it appears that, in a sense, although all the massless free fields do have zero rest mass, *only* the basic free fields propagate only along null directions. It is probably best to regard the nonbasic massless free fields simply as derivatives of basic free fields.

The appropriate analog of (4.6) for a general massless free field $\phi_A^{R'\cdots Y'}{}_{\cdots EF\cdots M} = \partial_F{}^{R'}\cdots\partial_M{}^{Y'}\theta_{A\dots E}$, say, with k primed indices and $n+k$ unprimed indices, its null datum being $\phi = D^k\theta$, can be obtained from the kth derivative of (4.6):

$$\partial_F{}^{R'}\cdots\partial_M{}^{Y'}\theta_{A\dots E}(P) = \frac{1}{2\pi}\int_{\mathcal{S}}\sum_{j=0}^{k}\frac{(-1)^{k-j}(n+k)!\,k!}{r^{k-j+1}(n+j)!\,(k-j)!\,j!}$$
$$\cdot\zeta_{(A}\cdots\zeta_J\xi_K\cdots\xi_{M)}\bar{\xi}^{(R'}\cdots\bar{\xi}^{V'}\bar{\xi}^{W'}\cdots\bar{\xi}^{Y')}\{D^{j+1}\theta - (n+2j+1)$$
$$\cdot\rho D^j\theta + j(n+j)(\rho^2 - \sigma\bar{\sigma})D^{j-1}\theta\}\,d\mathcal{S} \qquad (4.21)$$

In this expression it is supposed that there are $n+j$ ζ's, $k-j$ ξ's, j $\bar{\zeta}$'s, and $k-j$ $\bar{\xi}$'s.

The second and most difficult stage of the proof is to verify that the calculated null datum (on $\mathcal{N}^+$ and $\mathcal{N}^-$) for the basic free field $\theta_{A\dots E}(P)$ obtained using (4.6) is, in fact, the given function θ. (It will still be assumed, here, that the given field vanishes in the neighborhood of $\mathcal{B}$.) Since (4.6) is defined only for points P *not* lying on $\mathcal{N}$, in order to compare the calculated field with the given function θ, it is necessary to carry out a limiting procedure as P approaches a point O on $\mathcal{N}$ (see Figure 7).

Choose O (on $\mathcal{N}$) as origin of coordinates and let **OP** be a null vector. The point O is then a particular position of Q for P and there are particular positions $\xi^A(O)$, $\zeta^A(O)$ of ξ^A, ζ^A, respectively, associated with O. Choose a spinor reference frame so that

$$\xi_A(O) = \binom{0}{1}, \quad \zeta_A(O) = \binom{1}{0}, \quad \xi^A(O) = \binom{1}{0}, \quad \zeta^A(O) = \binom{0}{-1} \qquad (4.22)$$

This is consistent with (4.7) [and (1.3)].

Let $\mathcal{P}$ be the null hyperplane through OP. For a general position of Q, the generator of $\mathcal{N}$ through Q will meet $\mathcal{P}$ in a point S, say. The position spinor $s^{AB'}$ of S (with respect to O) is of the form

$$s^{AB'} = \begin{pmatrix} 0 & z \\ z & f \end{pmatrix} \qquad (4.23)$$

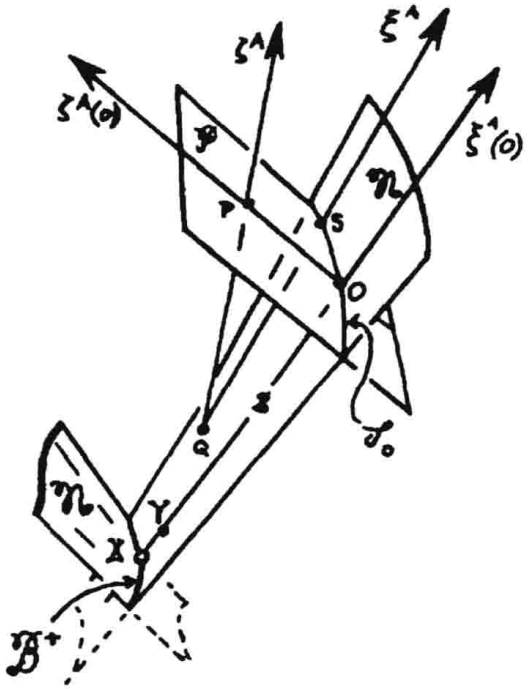


Fig. 7. As P approaches 0, the region of integration for (4.6) approaches a segment of a generator g of $\mathfrak{N}$. The determination of the field at neighboring points to $\mathfrak{N}$ from the null datum on $\mathfrak{N}$ is thus, in a sense, a nonlocal process and becomes somewhat complicated.

since it is orthogonal to the position spinor $p^{AB'}$ of P; this being *normal* to $\mathcal{P}$ and of the form

$$p^{AB'} = \epsilon \zeta^A(O)\bar{\zeta}^{B'}(O) = \begin{pmatrix} 0 & 0 \\ 0 & \epsilon \end{pmatrix} \tag{4.24}$$

(The scalar product $a_{AB'}b^{AB'}$ of two spinors $a^{AB'}$, $b^{AB'}$ is, of course, $a^{00'}b^{11'} + a^{11'}b^{00'} - a^{01'}b^{10'} - a^{10'}b^{01'}$.) The complex number z can be used to parametrize the two-dimensional intersection $\mathfrak{S}_0$ of $\mathcal{P}$ with $\mathfrak{N}$. It also parametrizes the set of generators of $\mathfrak{N}$. The real number f is then a function of z and $\bar{z}$. The real number ϵ is, of course, independent of z. It will ultimately be made to tend to zero.

Because of the freedom $\xi^A \to \lambda\xi^A$, we can choose

$$\xi^A = \begin{pmatrix} 1 \\ \xi \end{pmatrix} \tag{4.25}$$

The spinor equivalent of SQ then has the form

$$l\xi^A\bar{\xi}^{B'} = \begin{pmatrix} l & l\bar{\xi} \\ l\xi & l\xi\bar{\xi} \end{pmatrix} \tag{4.26}$$

Now, **SQ** is normal to $\mathcal{N}$ at S and is therefore also normal to $\mathcal{S}_0$, at S. The tangent plane of $\mathcal{S}_0$ at S is spanned by

$$\begin{pmatrix} 0 & 0 \\ 1 & \dfrac{\partial f}{\partial z} \end{pmatrix} \quad \begin{pmatrix} 0 & 1 \\ 0 & \dfrac{\partial f}{\partial \bar{z}} \end{pmatrix}$$

[see (4.23)], so the orthogonality with (4.26) gives

$$\frac{\partial f}{\partial z} = \bar{\xi}, \quad \frac{\partial f}{\partial \bar{z}} = \xi \tag{4.27}$$

Since $\zeta_A \xi^A = 1$, the spinor ζ^A has the form

$$\zeta^A = \begin{pmatrix} \zeta \\ \zeta\xi - 1 \end{pmatrix} \tag{4.28}$$

by (4.25). The spinor equivalent of **QP** is then

$$r\zeta^A \bar{\xi}^{B'} = \begin{pmatrix} r\zeta\bar{\zeta} & r\zeta(\bar{\zeta}\bar{\xi} - 1) \\ r\bar{\zeta}(\zeta\xi - 1) & r(\zeta\xi - 1)(\bar{\zeta}\bar{\xi} - 1) \end{pmatrix} \tag{4.29}$$

Now, **OS** + **SQ** + **QP** = **OP**, so from (4.23), (4.26), (4.29), and (4.24) we get

$$l + r\zeta\bar{\zeta} = 0 \tag{4.30}$$

$$z + l\xi + r\bar{\zeta}(\zeta\xi - 1) = 0 \tag{4.31}$$

$$f + l\xi\bar{\xi} + r(\zeta\xi - 1)(\bar{\zeta}\bar{\xi} - 1) = \epsilon \tag{4.32}$$

From (4.30) and (4.31) it follows that

$$\zeta = -\frac{l}{z} \tag{4.33}$$

$$l = -\frac{z\bar{z}}{r} \tag{4.34}$$

whence (4.32) gives

$$r = \epsilon - f + z\bar{\xi} + \bar{z}\xi \tag{4.35}$$

It is required to evaluate the limit, as P approaches O along PO (i.e., $\epsilon \to 0$), of $\theta_{AB\ldots E}(P)\zeta^A(O)\zeta^B(O)\cdots\zeta^E(O)$, where $\theta_{AB\ldots E}(P)$ is given by (4.6). In the limit, the region of integration $\mathcal{S}$ becomes the finite segment of the generator g of $\mathcal{N}$ through O, whose end points are O and the intersection X of g with the relevant $\mathcal{B}^+$ or $\mathcal{B}^-$. (This supposes P approaches $\mathcal{N}$ from one of $\mathcal{M}^+$, $\mathcal{M}^-$. If P were to approach $\mathcal{N}$ from $\mathcal{M}^0$, then the limit of $\mathcal{S}$ would consist of two portions of g extending to infinity.) Now, by assumption, the integrand (4.6) vanishes in

the neighborhood of X so the part of $\mathcal{S}$ lying in this neighborhood gives no contribution to (4.6). Call the remaining part $\mathcal{S}_1$. Then $\mathcal{S}_1$ converges on a proper subsegment OY of OX, where θ vanishes at Y. The segment OY will lie in a nonsingular part of $\mathcal{N}$, so the generators of $\mathcal{N}$ which intercept $\mathcal{S}_1$ must ultimately converge on g. Thus as $\epsilon \to 0$, the values of z which are of relevance to (4.6) must ultimately tend to zero. We may therefore assume that $|z| < k$ where $k \to 0$ as $\epsilon \to 0$.

Now using the notation of (3.14) but with $\xi^A(O)$ for ξ^A and $-\zeta^A(O)$ for η^A [(3.10) is satisfied and the definitions of z agree], we get

$$\xi^A = \xi^A(S) = \xi^A(O) + s^{BC'}[\partial_{BC'}\xi^A](O) + O(k^2)$$
$$= \xi^A(O) + \bar{z}[\delta\xi^A](O) + z[\bar{\delta}\xi^A](O) + O(k^2)$$

[see (4.23); f is clearly $O(k^2)$]. Contraction with $\xi^A(O)$ gives

$$\xi_A(O)\xi^A = \bar{z}[\xi_A\delta\xi^A](O) + z[\xi_A\bar{\delta}\xi^A](O) + O(k^2)$$

so by (4.22), (4.25), (3.16),

$$\xi = -\bar{z}\sigma_0 - z\rho_0 + O(k^2) \tag{4.36}$$

where σ_0, ρ_0 are the values of σ, ρ at 0. By (4.27) we have

$$f = -\tfrac{1}{2}\bar{\sigma}_0 z^2 - \rho_0 z\bar{z} - \tfrac{1}{2}\sigma_0\bar{z}^2 + O(k^3) \tag{4.37}$$

since $f = 0$ at 0.

Since ρ_0 and σ_0 differ from the values of ρ, σ at S by small quantities of order k, the value of ρ at Q is given by

$$\rho = \frac{\rho_0 + l(\sigma_0\bar{\sigma}_0 - \rho^2)}{(1 - l\rho_0)^2 - l^2\sigma_0\bar{\sigma}_0} + O(k) \tag{4.38}$$

[See (3.25); and the l parameters agree by (3.24), (4.26).] Also if

$$d\mathcal{S} = a\, d\mathcal{S}_0 \tag{4.39}$$

where $d\mathcal{S}, d\mathcal{S}_0$ are elements of surface area cut off by the same tube of generators of $\mathcal{N}$, on $\mathcal{S}$ and $\mathcal{S}_0$, respectively, then by (3.21),

$$\frac{\partial a}{\partial l} = -2\rho a \tag{4.40}$$

We have $a = 1$ at $l = 0$, whence by (4.38),

$$a = (1 - l\rho_0)^2 - l^2\sigma_0\bar{\sigma}_0 + O(k) \tag{4.41}$$

Let

$$z = \frac{ue^{i\omega}}{2^{1/2}} \tag{4.42}$$

where u, ω are real parameters, giving polar coordinates for the surface $\mathcal{S}_0$ [u being the distance OS and ω an angular parameter for S; see (4.23)]. The required component of the integral (4.6) may now be written

$$\Theta = \theta_{AB\cdots E}(P)\,\xi^A(O)\,\xi^B(O)\cdots\xi^E(O)$$

$$= \frac{1}{2\pi}\int_0^\infty \int_0^{2\pi} \frac{1}{r}(1-\xi\zeta)^n a^{-(n-1)/2}\,\frac{\partial}{\partial l}\{a^{(n+1)/2}\theta\}\cdot au\,d\omega\,du \quad (4.43)$$

by (4.40), (4.39) since by (4.28), (4.22) $\zeta_A\xi^A(0) = (1-\xi\zeta)$. The integrand is to be taken as zero outside the relevant region corresponding to $\mathcal{S}_1$. It will be convenient to rewrite (4.43), however, as an integral over w and l, where

$$w = e^{-i\omega} \quad (4.44)$$

so that

$$dw = -iwd\omega$$

$$w^2 = \frac{\bar{z}}{z}, \qquad u^2 = 2z\bar{z}$$

From (4.33), (4.36) we have

$$1-\xi\zeta = 1 - l\rho_0 - w^2 l\sigma_0 + O(k) \quad (4.45)$$

From (4.35), (4.36), (4.37)

$$r = \epsilon - \frac{u^2}{4w^2}(\bar{\sigma}_0 + 2\rho_0 w^2 + \sigma_0 w^4) + O(k^3)$$

which with (4.34) leads to

$$dl = \frac{u}{2rw^2}\{l\bar{\sigma}_0 + 2(l\rho_0 - 1)w^2 + l\sigma_0 w^4 + O(k)\}\,du + \{\cdots\}\,dw \quad (4.46)$$

Thus substituting (4.41), (4.44), (4.45), and (4.46) into (4.43), we get

$$\Theta = \frac{1}{2\pi}\int_0^{l_1}(A^2 - B\bar{B})^{(1-n)/2}\,\frac{d}{dl}\{(A^2 - B\bar{B})^{(n+1)/2}\theta\}$$

$$\left\{-\oint \frac{2iw(A - Bw^2)^n}{(\bar{B} - 2Aw^2 + Bw^4)}\,dw\right\}\,dl + O(k) \quad (4.47)$$

where

$$A = 1 - l\rho_0, \qquad B = l\sigma_0$$

and l_1 is the value of l at Y. The contour for the w integral is the unit circle. The values of θ to be considered here are those associated with the points of

OY. These differ from the corresponding values on $\mathcal{S}_1$ by quantities of order k (assuming suitable smoothness of the function θ). Now the roots of $\bar{B} - 2Aw^2 + Bw^4$ (assuming $B \neq 0$) are given by

$$w^2 = \frac{A \pm (A^2 - B\bar{B})^{1/2}}{B}$$

Two of these, w_1, w_2 (given by a minus sign above) lie inside the unit circle and the other two lie outside the unit circle. This is because $A^2 - B\bar{B} = a$ on OY with $a > 0$ (and $A > 0$), $\mathcal{N}$ being nonsingular along OY, so $A > |B|$. Hence,

$$\oint \frac{2iw(A - Bw^2)^n}{(\bar{B} - 2Aw^2 + Bw^4)}\, dw = 2\pi i \left\{ \frac{2iw_1(A - Bw_1{}^2)^n}{-4Aw_1 + 4Bw_1{}^3} + \frac{2iw_2(A - Bw_2{}^2)^n}{-4Aw_2 + 4Bw_2{}^3} \right\}$$
$$= 2\pi(A^2 - B\bar{B})^{(n-1)/2}$$

If $B = 0$ this is clearly still true. Thus, (4.47) gives

$$\Theta = -[(A^2 - B\bar{B})^{(n+1)/2}\theta]_{l=0}^{l=l_1} + O(k)$$

But $\theta = 0$ at $l = l_1$, and $A = 1, B = 0$ at $l = 0$. Therefore, the limiting value of Θ as P approaches O is, in fact, the value of θ at O as required.

The above argument can also be carried through, under reasonable assumptions, if P approaches O from $\mathcal{M}^0$, except that at the last stage there will be an extra contribution to Θ unless for each generation of $\mathcal{N}$

$$\mathcal{Q}\,[l^{n+1}\theta]_{-\infty}^{+\infty} = 0 \tag{4.48}$$

This is the condition of "continuity at infinity" that has been alluded to earlier. It is to be assumed that both limits exist, which implies

$$\theta = O(l^{-n-1}) \tag{4.49}$$

both for large positive l and large negative l. Condition (4.49) appears to be an important one in the radiation theory[24] for $\theta_{AB\ldots E}$. If it is assumed that both limits in (4.48) exist but that they are not necessarily equal, then it appears to be possible modify (4.6) and (4.9) so that they will remain true for points P in $\mathcal{M}^0$. To do this, a contribution corresponding to each null line b through P which "touches $\mathcal{N}$ at infinity" must be added. Such a line b has the property that the null hyperplane through it also contains a generator h of $\mathcal{N}$. Let $h^{AB'}$ be the spinor equivalent of **PH**, where H is a point on h. Let ξ^A and l be associated with h in the usual way (i.e., ξ^A is constant along h and $l\xi^A\bar{\xi}^{B'}$ gives the posi-

[24] This condition can be expected to hold if both the incoming and outgoing radiation have a sufficiently curtailed duration. In the gravitational case, the condition $\psi = 0(l^{-5})$ has been used [14] (in *curved* space) as a basis for a study of the asymptotic behavior of ψ_{ABCD}.

 Null hypersurface initial data for classical fields of arbitrary spin and for general relativity. Widely circulated preprint, *Cornell* (1961). Also circulated in P. G. Bergmann's ARL Technical Documentary Report (1963) 63–56; *Gen. Relativity Gravitation* **12** (1980), no. 3, 225–64

tion spinor of a general point on h). Then the expression to be added to the right-hand side of (4.6) or (4.9) appears to be

$$\frac{2^{n-1}}{\pi}\int_{\mathcal{I}}[I^{n+1}\theta]_{-\infty}^{+\infty}(-h_{FW'}h^{FW'})^{-n-1/2}h_{AR'}\bar{\xi}^{R'}h_{BS'}\bar{\xi}^{S'}\cdots h_{EV'}\bar{\xi}^{V'}\,d\mathcal{I} \quad (4.50)$$

where $d\mathcal{I}$ is the element of proper arc length of the curve (or series of curves) traced out by the point whose position spinor relative to some fixed origin is $\xi^A\bar{\xi}^{B'}$. The spinor $h_{AR'}\bar{\xi}^{R'}$ is in fact proportional to ξ_A and both it and $h_{FW'}h^{FW'}$ ($= -PH^2$) are independent of the choice of H on h. (Also, it is not hard to verify that (4.50) is invariant under $\xi^A \to \lambda\xi^A$.) If (4.50) is added to (4.6) then it can be shown that the above limiting argument will hold also for points P in $\mathfrak{M}^0$

For the uniqueness proof, it will not be needed to assume that the given data are zero in the neighborhood of $\mathcal{B}$. However, this makes little difference in fact. Once (4.6) has been established under the assumption of *vanishing* data in the neighborhood of $\mathcal{B}$, this assumption can readily be dispensed with, provided of course that the null cone $\mathcal{C}$ of P does not meet $\mathcal{B}$. For if $\mathcal{C}$ does not meet $\mathcal{B}$, then it does not meet some neighborhood of $\mathcal{B}$ (since $\mathcal{B}$ is a closed set). But, because of the form of (4.6), and the related Green's functions, the field at P can only be influenced[25] by data that happen to lie on $\mathcal{C}$. Therefore there can be no contribution to the field at P from data in such a neighborhood of $\mathcal{B}$. Moreover if $\mathcal{C}$ *does* meet $\mathcal{B}$ the correct formula [see (4.6)] can be obtained by a limiting procedure, as was described earlier.

As a first stage in the uniqueness proof, it will be shown that the field on $\mathcal{N}^+$ is determined by the field on $\mathcal{B}^+$ and the null datum on $\mathcal{N}^+$ (and similarly for $\mathcal{N}^-$, $\mathcal{B}^-$). The argument will be carried though for a general massless free field $\theta_{A\cdots H}^{P'\cdots V'}$ with m primed and n unprimed indices.

Define, with p η's and q $\bar{\eta}$'s, components

$$\theta_{(p,q)} = \theta_{A\cdots CD\cdots H}^{P'\cdots S'T'\cdots V'}\,\bar{\eta}_{P'}\cdots\bar{\eta}_{S'}\bar{\xi}_{T'}\cdots\bar{\xi}_{V'}\eta^A\cdots\eta^C\xi^D\cdots\xi^H \quad (4.51)$$

where ξ^A and η^A are as in Section 3, that is, ξ^A is associated with $\mathcal{N}$ in the usual way and η^A is such that it is constant along each generator of $\mathcal{N}$, with $\xi_A\eta^A = 1$. Let us, in fact, choose η^A to be constant everywhere. Then this fixes the normalization of ξ^A on each generator except for the particular generator g^* which happens to be parallel to the direction of η^A. Let the operators $D, \delta, \bar{\delta}$ be as in (3.13), (3.14); then

$$D\xi^A = 0, \quad D\eta^A = 0, \quad \delta\eta^A = 0, \quad \bar{\delta}\eta^A = 0 \quad (4.52)$$

[25] This relates to the fact that if the field is zero throughout an appropriate region of $\mathcal{C}$ which meets every generator of $\mathcal{C}$, then it is zero at P. This can also be proved from the fact that the field satisfies D'Alembert's equation.

Also, by (3.16) $\bar{\delta}\xi^A$ differs from $-\rho\eta^A$ by a multiple of ξ^A. This multiple must be zero since $\bar{\delta}(\eta_A \xi^A) = \eta_A \bar{\delta}\xi^A = 0$. Thus,

$$\bar{\delta}\xi^A = -\rho\eta^A, \qquad \delta\bar{\xi}_{P'} = -\rho\bar{\eta}_{P'} \tag{4.53}$$

and similarly

$$\delta\xi^A = -\sigma\eta^A, \qquad \bar{\delta}\bar{\xi}_{P'} = -\sigma\bar{\eta}_{P'} \tag{4.54}$$

Applying (4.1) to the D derivative of (4.51) and using (4.52), (4.53), (4.54) we obtain

$$\{D - (m+1-p)\rho\}\,\theta_{(p,q)} = \delta\theta_{(p-1,q)} + (n-q)\,\sigma\theta_{(p-1,q+1)} \tag{4.55}$$

$$\{D - (n+1-q)\rho\}\,\theta_{(p,q)} = \bar{\delta}\theta_{(p,q-1)} + (m-p)\,\bar{\sigma}\theta_{(p+1,q-1)} \tag{4.56}$$

Assuming one of the right-hand sides to be a known quantity on a segment g' of a generator $g(\neq g^*)$ of $\mathcal{N}$, this gives us an ordinary differential equation of a standard type, which can be used to determine $\theta_{(p,q)}$ uniquely along g', provided $\theta_{(p,q)}$ is known at some initial point B on g'. (Of course, to deal with g^* a different choice of η^A must be made.) If $m = 0 = p$ (so that the field is a basic free field), then (4.56) enables us to calculate $\theta_{(0,1)}, \theta_{(0,2)}, \ldots, \theta_{(0,n)}$ successively from θ $(=\theta_{(0,0)})$[26] and known initial values of the field $\theta_{A\ldots H}$. Since (4.56) involves the tangential derivative $\bar{\delta}$, θ must be given on more than just a segment of a single generator g, but it will suffice to give θ on a portion of $\mathcal{S}$ containing a set of neighboring segments of generators to g', and to give $\theta_{A\ldots H}$ at one point on each of these segments. Hence, knowledge of the field on $\mathcal{B}^+$ and the null datum on $\mathcal{N}^+$ is sufficient for the determination the field on $\mathcal{N}^+$. Similarly, the corresponding result holds for $\mathcal{B}^-$, $\mathcal{N}^-$.

If $m \neq 0 \neq n$, then (4.53) and (4.54) may be used successively in some order to obtain all the $\theta_{(p,q)}$'s if θ and initial values of $\theta_{A\ldots H}^{P'\ldots V'}$ are known. Thus, for *all* the massless free fields, the field on $\mathcal{B}^+$ and the null datum on $\mathcal{N}^+$ together determine the field on $\mathcal{N}^+$ as above (and similarly for $\mathcal{B}^-$, $\mathcal{N}^-$). However, for the nonbasic fields there is a redundancy implied by (4.55), (4.56). For if we take the difference between these two equations we get a relation connecting the $\theta_{(p,q)}$'s not involving their D derivatives. By a suitable choice of η^A it can, in general, be arranged that at a point X in $\mathcal{B}$, the derivatives δ, $\bar{\delta}$ operate in the tangent plane to $\mathcal{B}$. This implies that there are differential relations connecting the components of the field $\theta_{A\ldots H}^{P'\ldots V'}$ on $\mathcal{B}$, so that the initial data cannot, in general, be chosen arbitrarily on $\mathcal{B}$. On the other hand, in the case of the *basic* free fields this difficulty does not arise and the field on $\mathcal{B}$ can, in fact, be chosen arbitrarily (and independently of the null datum on $\mathcal{N}^+$ and $\mathcal{N}^-$). This can be established by examining in detail the $n+1$ independent Green's functions that can be obtained for (4.9), (4.10), (4.11) corresponding to the various components of the field at a given point of $\mathcal{B}$.

[26] It is possible to carry through a similar argument in general relativity also [14].

 Null hypersurface initial data for classical fields of arbitrary spin and for general relativity. Widely circulated preprint, *Cornell* (1961). Also circulated in P. G. Bergmann's ARL Technical Documentary Report (1963) 63–56; *Gen. Relativity Gravitation* **12** (1980), no. 3, 225–64

For the remainder of the uniqueness proof it will be convenient to introduce, for each massless free field $\theta_{AB\cdots E}^{U'\cdots W'}$, a certain expression $T_{\alpha\beta\cdots\epsilon:\rho\cdots\upsilon}$, which will be termed the *field density tensor.* Its spinor equivalent is given by

$$T_{AF'BG'\cdots EJ':RV'\cdots UY'} = \theta_{AB\cdots EV'\cdots Y'}\,\bar{\theta}_{F'G'\cdots J'R\cdots U} \tag{4.57}$$

It is readily verified from (4.57) [using (4.6)] that the field density tensor has vanishing divergence on each index:

$$\partial^{\alpha}T_{\alpha\beta\cdots\epsilon:\rho\cdots\upsilon} = 0, \qquad \partial^{\rho}T_{\alpha\beta\cdots\epsilon:\rho\sigma\cdots\upsilon} = 0 \tag{4.58}$$

complete symmetry in each of its two groups of indices:

$$T_{\alpha\beta\cdots\epsilon:\rho\cdots\upsilon} = T_{(\alpha\beta\cdots\epsilon):(\rho\cdots\upsilon)}$$

and vanishing traces within each group of indices:

$$T^{\beta}{}_{\beta\gamma\cdots\epsilon:\rho\sigma\cdots\upsilon} = 0, \qquad T_{\alpha\beta\cdots\epsilon:}{}^{\sigma}{}_{\sigma\cdots\upsilon} = 0$$

Of course, for a basic free field one of the groups of indices will be vacuous.

The field density tensor for the Maxwell field is simply (half) the energy-momentum tensor. In the case of a gravitational field (given by ψ_{ABCD}) is is proportional to the "gravitational density" (or "superenergy") tensor of Robinson and Bel [27] (see also [13]) although the above definition was, strictly speaking, only given in flat space. For the Weyl neutrino field it gives the "current vector" [10]. In the case of a D'Alembert scalar field χ it gives simply the square of the modulus of the field. However, the energy-momentum tensor for χ can be obtained as the field density tensor (symmetrized if χ is complex) of the gradient field $\partial_A{}^{P'}\chi$ of χ.

The field density tensors have the further property of positive definiteness:

$$T_{\alpha\beta\cdots\epsilon:\rho\cdots\upsilon}a^{\alpha}b^{\beta}\cdots e^{\epsilon}r^{\rho}\cdots u^{\upsilon} \geqslant 0 \tag{4.59}$$

for any set of vectors $a^{\alpha}, b^{\beta}, \ldots, e^{\epsilon}, r^{\rho}, \ldots, u^{\upsilon}$, each of which is either timelike or null and all of which point into the future. If *all* these vectors are timelike then the *strict* inequality holds unless $\theta_{AB\cdots E}^{R'\cdots U'} = 0$.

To prove (4.59), we observe first that it is true if all of $a^{\alpha}, \ldots, u^{\upsilon}$ are null and pointing into the future, for then their spinor equivalents $a^{AR'}, \ldots, u^{AR'}$ have the form $\alpha^{A}\bar{\alpha}^{R'}, \ldots, \upsilon^{A}\bar{\upsilon}^{R'}$ and so by (4.57) the left-hand side of (4.59) is of the form

$$|\theta_{AB\cdots E}^{R'\cdots U'}\alpha^{A}\beta^{B}\cdots\epsilon^{E}\bar{\rho}_{R'}\cdots\bar{\upsilon}_{U'}|^2 \tag{4.60}$$

Now any future-pointing timelike vector can be expressed (in infinitely many ways) as a sum of two future-pointing null vectors. Hence, the left-hand side of (4.59) can be expressed as a sum of expressions like (4.60) and so must be positive on zero. If all of $a^{\alpha}, \ldots, u^{\upsilon}$ are timelike, we can ensure that at least one of the terms of the sum be strictly positive unless $\theta_{AB\cdots E}^{R'\cdots U'} = 0$ because of the freedom of representing the timelike vectors as sums of null vectors.

The uniqueness proof will now be completed.[27] It will be shown that if the field on $\mathfrak{N}^+$ (and $\mathfrak{B}^+$) is known then the field in $\mathfrak{M}^+$ is completely determined. The argument for $\mathfrak{M}^-$ is similar. It does not seem, however, that this type of argument can be directly applied to obtain uniqueness in $\mathfrak{M}^0$ (but it does appear to be possible to obtain the result for $\mathfrak{M}^0$ from that for $\mathfrak{M}^+$ by means of an argument involving a conformal transformation of the space-time).

Suppose that the field on $\mathfrak{N}^+$ and $\mathfrak{B}^+$ does not determine the field at some point P in $\mathfrak{M}^+$. Then there will be two solutions of the field equations which differ at P but are identical on $\mathfrak{N}^+$ and $\mathfrak{B}^+$. Their difference will therefore satisfy the field equations, be zero on $\mathfrak{N}^+$ and $\mathfrak{B}^+$, but be nonzero at P. It is required to show that this is impossible.

Choose a future-pointing constant timelike vector v^μ. Then by (4.58), the vector

$$w_\alpha = T_{\alpha\beta\cdots\epsilon:\rho\cdots\upsilon} v^\beta \cdots v^\epsilon v^\rho \cdots v^\upsilon$$

has vanishing divergence and $v^\alpha w_\alpha > 0$ unless $\theta^{R'\cdots U'}_{AB\cdots E} = 0$. Consider the closed hypersurface which is made up of that part $\mathfrak{N}_1$ of $\mathfrak{N}^+$ and $\mathfrak{B}^+$ that lies in the interior of the past null cone $\mathcal{C}$ of P and of the part $\mathcal{C}_1$, of $\mathcal{C}$ which connects the boundary of this region to P. The vector w_α vanishes on $\mathfrak{N}_1$ whence by Gauss's theorem the integral over $\mathcal{C}_1$ (suitably expressed) of the scalar product of w_α with the "outward drawn" normal to $\mathcal{C}_1$ vanishes. But this normal is a future-pointing null vector C^α, whence by (4.59) its scalar product $C^\alpha w_\alpha$ with w_α must be positive or zero. Since the integral of $C^\alpha w_\alpha$ over $\mathcal{C}_1$ is zero this implies (assuming continuity) that $C^\alpha w_\alpha = 0$ everywhere on $\mathcal{C}_1$. In particular this must hold at the vertex P, and since P may be approached on $\mathcal{C}_1$ in *any* null direction this implies $w_\alpha = 0$ at P. Hence $\theta^{R'\cdots U'}_{AB\cdots E} = 0$ at P as it was required to show.

This completes the proof for all cases except the D'Alembert field. In that case there are no indices available with which to construct the w_α vector. However, this difficulty can easily be circumvented by applying the argument instead to the gradient of the D'Alembert field (see also [4]). In fact the uniqueness for the D'Alembert case, once established, could have been used to help establish uniqueness in all the other cases since all the massless free fields satisfy D'Alembert's equation. The introduction of all the various field density expressions was therefore not strictly necessary, but it is felt that these expressions have sufficient interest and importance in themselves to warrant the use of them here.

If t^μ is a unit future-pointing timelike vector, the "field density"

$$T_{\alpha\beta\cdots\upsilon} t^\alpha t^\beta \cdots t^\upsilon \tag{4.61}$$

is positive unless $\theta^{R'\cdots U'}_{AB\cdots E} = 0$. It is interesting to observe that if t^μ is replaced

[27] I am indebted to E. Schuecking for part of the idea of this proof.

by the null vector corresponding to $\xi^A \bar{\xi}^{B'}$, then (4.61) is simply the square of the modulus of the null datum associated with ξ^A. In fact, $|\theta|^2$ may be regarded as a measure of the "flux of field density" across $\mathfrak{N}$.

The remaining information contained in θ, namely, the phase angle, has a directional interpretation related to polarization. This interpretation depends on the spin of the massless free field $\theta_{AB\ldots E}^{R'\cdots U'}$.. Supposing that there are p primed and q unprimed indices, we saw in Section 3 that there were $\frac{1}{2}|p-q|$ positions of the "flag" corresponding to ξ^A which gave the same value to θ. That is to say, the geometrical interpretation of θ has a $\frac{1}{2}|p-q|$-fold axis of symmetry about the "flagpole" of ξ^A. If $p=q$ this is to be interpreted as a complete axial symmetry. If $p-q$ is odd the interpretation involves the two-valuedness of $\theta_{A\cdots E}^{R'\cdots U'}$.

This is a local symmetry given by θ, but it is clear that global symmetries[28] possessed by solutions of the field equations may exhibit similar local symmetries. For example, any spherically symmetric solution is given by a spherically symmetric null datum on a null cone. Unless $p=q$ (i.e., unless the spin is zero), this null datum must vanish. This essentially expresses the well-known fact that monopole radiation (for a massless free field) must have zero spin. Corresponding results for higher multipoles can also be obtained in this way.

Acknowledgments

I am grateful to my colleagues at Princeton, Syracuse, and Cornell for many helpful discussions, and to C. W. Misner, I. Robinson, E. Schuecking, A. Trautman, A. Komar, E. T. Newman, and D. W. Sciama, in particular. I also received valuable stimulation in discussions with O. Penrose and with R. K. Sachs.

Note

Since this article was written in 1961, a great number of improvements in technique and formalism have taken place. I have refrained from making any changes, however, except for the correction of some minor errors. For an argument leading to a more streamlined proof of the result of §(4), see [19].

[28]There is also a curious global topological property of the null datum which is implied by these local symmetries, if continuity is assumed. This is that every spacelike section of $\mathfrak{N}^+$ (not meeting $\mathfrak{B}$) must contain a net algebraic count of $q-p$ places where θ vanishes. These places will, in general, describe at least $|q-p|$ lines of zeros of θ on $\mathfrak{N}^+$, and in the case of a cone, there will be a net count of $q-p$ of these zero lines entering the vertex. The directions of these zero lines at the vertex define *principal null directions* of the field at the vertex. If $p=0$ there will, in general, be exactly q such directions (see also [13]).

References

1. Lichnerowicz, A. (1955). *Théories relativistes de la gravitation et de l'électromagnétisme* (Masson, Paris).
2. Foures-Bruhat, Y. (1956). *J. Rat. Mech. Anal.*, **5**, 941.
3. Bondi, H., van der Burg, M. G. J., and Metzner, A. W. K. (1962). *Proc. R. Soc. London Ser. A*, **269**, 21; Sachs, R. K. (1962). *Proc. R. Soc. London Ser. A*, **270**, 103.
4. Courant, R., and Hilbert, D. (1937). *Methoden der Mathematischen Physik I, II* (Interscience, New York).
5. Riesz, M. (1949). *Acta Math.*, **81**, 1.
6. Infeld, L., and van der Waerden, B. L. (1933). *Sitz. Ber. Preuss. Acad., Wiss. Physik.-math, Kl.*, **9**, 380.
7. Whittaker, E. T. (1937). *Proc. R. Soc. London Ser. A*, **158**, 38.
8. Ruse, H. S. (1944). *Proc. R. Soc. Edinburg A*, **62**, 64.
9. Bade, W. L., and Jehle, H. (1953). *Rev. Mod. Phys.*, **25**, 714.
10. Corson, E. M. (1953). *Introduction to Tensors, Spinors, and Relativistic Wave-equations* (Blackie, London).
11. Veblen, O., and von Neumann, J. (1935–36). *Princeton Institute for Advanced Study-Lectures*, "Geometry of Complex Domains" (Princeton University Press, Princeton, New Jersey).
12. Payne, W. T. (1959). *Am. J. Phys.*, **27**, 318.
13. Penrose, R. (1960). *Ann. Phys. (N.Y.)*, **10**, 171.
14. Newman, E. T., and Penrose, R. (1962). *J. Math. Phys.*, **3**, 566; **4**, 998.
15. Schrödinger, E. (1950). *Space-Time Structure* (Cambridge University Press, Cambridge).
16. Witten, L. (1959). *Phys. Rev.*, **113**, 357.
17. Dirac, P. A. M. (1936). *Proc. R. Soc. London Ser. A*, **155**, 447.
18. Hadamard, J. (1952). *Lectures on Cauchy's Problem in Linear Partial Differential Equations* (Dover, New York).
19. Newman, E. T., and Penrose, R. (1968). *Proc. R. Soc. London Ser. A*, **305**, 175.
20. Robinson, I. (1961). *J. Math. Phys.*, **2**, 290
21. Sachs, R. K. (1961). *Proc. R. Soc. London Ser. A*, **264**, 309.
22. Pirani, F. A. E. (1960). *ARL Technical Note* 60-143.
23. Penrose, R. (1966). In *Perspectives in Geometry and Relativity*, ed. Hoffmann, B. (Indiana University Press, Bloomington).
24. Penrose, R. (1972). In *General Relativity, Papers in Honour of J. L. Synge* ed. O'Raifeartaigh, L. (Clarendon Press, Oxford) Geroch, R., Held, A., and Penrose, R. (1973). *J. Math. Phys.*, **14**, 874.
25. Hlavatý, V. (1959). *J. Math. Mech.*, **8**, 285.
26. Wigner, E. P., and Bargmann, V. (1948). *Proc. Nat. Acad. Sci.*, **34**, 211.
27. Bel, L. (1959). *Compt. Rend.*, **248**, 1297.

115
THE COMPLEX GEOMETRY OF THE NATURAL WORLD

Proceedings of the International Congress of Mathematicians (1978), pp. 189–94, *Acad. Sci. Fennica, Helsinki*, 1980

This was the Plenary talk I gave at the International Congress of Mathematicians (ICM) in Helsinki 1978.

618 The complex geometry of the natural world. *Proceedings of the International Congress of Mathematicians* (1978), pp. 189–94, *Acad. Sci. Fennica, Helsinki,* 1980

Proceedings of the International Congress of Mathematicians
Helsinki, 1978

The Complex Geometry of the Natural World

Roger Penrose

This century has seen two major revolutions in physical thought. The first of these, relativity, uprooted earlier ideas of the nature of time and space, and provided us with our present picture of the world as a real differentiable manifold of dimension four, possessing a pseudo-Riemannian metric with a $(+\ -\ -\ -)$ signature. The second revolution, quantum theory, altered our picture of things yet more radically than did relativity—even to the extent that, as we were told, it became no longer appropriate to form pictures at all, in order to give accurate representations of physical processes on the quantum scale. And, for the first time, the complex field $\boldsymbol{C}$ was brought into physics at a fundamental and universal level, not just as a useful or elegant device, as had often been the case earlier for many applications of complex numbers to physics, but at the very basis of physical law. Thus, the allowable physical states were to form a complex vector space, in fact a Hilbert space. So, on the one hand, we had the real-manifold picture of space-time geometry, and on the other, the complex-vector space view, according to which geometrical pictures were deemed inappropriate.

This conflict has remained with us since the conceptions of these great theories, to the extent that, even now, there is no satisfactory union between the two. Even at the most elementary level, there are still severe conceptual problems in providing a satisfactory interpretation of quantum mechanical observations in a way compatible with the tenets of special relativity. And quantum field theory, which represents the fully special-relativistic version of quantum theory, though it has had some very remarkable and significant successes, remains beset with inconsistencies and divergent integrals whose illeffects have been only partially circumvented. Moreover, the present status of the unification of *general* relativity with quantum mechanics

remains merely a collection of hopes, ingenious ideas and massive but inconclusive calculations.

In view of this situation it is perhaps not unreasonable to search for a different viewpoint concerning the role of geometry in basic physics. Broadly speaking, "geometry", after all, means any branch of mathematics in which pictorial representations provide powerful aids to one's mathematical intuition. It is by no means necessary that these "pictures" should refer just to a spatio-temporal ordering of physical events in the familiar way. And since C plays such a basic universal role at the primitive levels of physics at which quantum phenomena are dominant, one is led to expect that the primitive geometry of physics might be complex rather than real. Moreover, the macroscopic geometry of relativity has many special features about it that are suggestive of a hidden complex-manifold origin, and of certain deep underlying physical connections between the normal spatio-temporal relations between things and the complex linear superposition of quantum mechanics.

One of the most striking and elementary of these physical connections appears in the geometry of the spin states of an ordinary spin $\frac{1}{2}\hbar$ particle (such as an electron or proton). Each pure state corresponds to a direction in space, about which we may regard the particle as spinning in a right-handed sense. But these states refer also to the rays through the origin of a 2-complex-dimensional Hilbert space. Thus we have an unambiguous correspondence between the complex directions of the Hilbert 2-space and the real directions of ordinary Euclidean 3-space. The set of complex Hilbert-space directions has a natural structure as a complex-projective (Kähler) 1-manifold $\boldsymbol{CP}^1$, while the real Euclidean directions constitute the ordinary 2-sphere S^2. The essential identity between these two spaces provides a link between, on the one hand, the fact that it is *complex* linear superposition that occurs in quantum mechanics and, on the other, that ordinary space is 3-dimensional! In terms of the symmetry groups involved, this relation finds expression in the special local isomorphism

$$SU(2) \to SO(3)$$

(which is a 2-1 map).

The group $SO(3)$ refers only to spatial symmetry. According to relativity, time must be brought in also, and the local symmetry about a point in space-time becomes the Lorentz group, whose connected component $O_+^\uparrow(1, 3)$, being a complex group, has an even more intimate relation to C than has $SO(3)$, the relativistic version of the above local isomorphism being

$$SL(2, \boldsymbol{C}) \to O_+^\uparrow(1, 3)$$

(which is again a 2-1 map). We may again think of these groups as acting on $\boldsymbol{CP}^1$ ($\cong S^2$), but where now we are concerned with *all* holomorphic (i.e. complex-analytic) maps of $\boldsymbol{CP}^1$ to itself (the Kähler structure being now dropped), giving precisely $O_+^\uparrow(1, 3)$. Indeed, we can (literally!) *see* this $\boldsymbol{CP}^1$ as the celestial sphere or "sky" of an observer [1]. Two observers who pass one another at the same event O,

 The complex geometry of the natural world. *Proceedings of the International Congress of Mathematicians* (1978), pp. 189–94, *Acad. Sci. Fennica, Helsinki,* 1980

have their skies related, according to the relativistic transformation law, by a conformal map, i.e. by a holomorphic map, if we regard the sky as $\boldsymbol{CP}^1$.

In space-time terms this $\boldsymbol{CP}^1$ is to be regarded as the set of light rays through the point O. Thus it is in the *light-ray* structure of space-time that complex geometry begins to emerge. (The relation to quantum superposition is, for the moment, less evident, but it reappears at a later stage.) Let us explore this further by considering the space $\mathscr{N}$ of light rays in ordinary Minkowski (i.e. flat pseudo-Euclidean $(+\ -\ -\ -)$ signature) 4-space M'. Since $\mathscr{N}$ is 5-real-dimensional it cannot be a complex manifold. Nevertheless, it is very close to being one and has the structure of what is known as a CR-hypersurface [2], [10]. This structure may be viewed as that which is induced on $\mathscr{N}$ by its imbedding as a real hypersurface in a complex manifold—and, in this case, the ambient complex manifold can, in fact [3], [10], be taken to be $\boldsymbol{CP}^3$. The tangent space D to a point of $\mathscr{N}$ can be thought of as the family of light rays that lie in the immediate neighbourhood of a given light ray. In D are certain real 2-planes, namely those, called *holomorphic directions,* that are spanned by the real and imaginary parts of complex vectors in $\boldsymbol{CP}^3$. Together, these sweep out a 4-real-dimensional or 2-complex-dimensional subspace of D, called the holomorphic tangent space. This represents light rays that are displaced from the given one by a small amount orthogonal to the direction of motion. A certain 1-real-parameter family of holomorphic directions arises from the system of light cones of the points of the given light ray. This family provides the directions along which the *Levi form* [2] (which measures the holomorphic convexity properties of the hypersurface $\mathscr{N}$ in $\boldsymbol{CP}^3$) vanishes. The Levi form eigenvalues for $\mathscr{N}$ have signs $(+\ -)$.

The space $\boldsymbol{CP}^3$, which in this context is referred to as the *projective twistor space* for M', can be constructed in the following way [3], [5], [6], [7]. Let $\boldsymbol{CM}'$ be the standard complexification of M', and let $(z^0, z^1, z^2, z^3) \in \boldsymbol{C}^4$ be standard coordinates for a point $z \in \boldsymbol{CM}'$, the complex metric being $ds^2 = (dz^0)^2 - (dz^1)^2 - (dz^2)^2 - (dz^3)^2$. Now consider the matrix equation

$$\begin{pmatrix} Z^0 \\ Z^1 \end{pmatrix} = \frac{i}{\sqrt{2}} \begin{pmatrix} z^0 + z^1 & z^2 + iz^3 \\ z^2 - iz^3 & z^0 - z^1 \end{pmatrix} \begin{pmatrix} Z^2 \\ Z^3 \end{pmatrix},$$

where the ratios $Z^0 : Z^1 : Z^2 : Z^3$ provide coordinates for the points of $\boldsymbol{CP}^3$. Holding these ratios fixed, the solutions of this equation give (provided that Z^2 and Z^3 do not both vanish) a complex-2-plane along which the induced metric vanishes. Such a plane is called *totally null,* and the particular ones arising, as above, from *points* of $\boldsymbol{CP}^3$ are called *α-planes.* There is also another family of totally null planes in $\boldsymbol{CM}'$, called *β*-planes. These arise in a corresponding way from the various *planes* ($\boldsymbol{CP}^2$'s) in $\boldsymbol{CP}^3$. If an α-plane (or a *β*-plane) has one real point in $\boldsymbol{CM}'$ (i.e., in M'), then it has a whole light ray of real points, the light ray then determining the α-plane (and its complex conjugate, the *β*-plane) uniquely. In this case the Hermitian form

$$2\Sigma(Z^\alpha) \equiv |Z^0 + Z^2|^2 + |Z^1 + Z^3|^2 - |Z^0 - Z^2|^2 - |Z^1 - Z^3|^2$$

vanishes. The equation $\Sigma(Z^\alpha)=0$, in $\boldsymbol{CP}^3$, defines $\mathcal{N}$ (which we take compact $\cong S^2\times S^3$).

Now suppose instead that, in the above matrix equation, z^α is held fixed and Z^α allowed to vary. Then we obtain a 2-complex-dimensional subspace of $\boldsymbol{C}^4$, i.e. a complex *line* ($\boldsymbol{CP}^1$) in $\boldsymbol{CP}^3$. Therefore, points of $\boldsymbol{CM}'$ are represented as lines in $\boldsymbol{CP}^3$. In fact the only lines of $\boldsymbol{CP}^3$ that do not so represent points of $\boldsymbol{CM}'$ are the lines meeting the particular line I in $\mathcal{N}$, given by $Z^2=Z^3=0$. To include these lines, we adjoin extra points to $\boldsymbol{CM}'$ to obtain a *compact* manifold $\boldsymbol{CM}$, this being the standard Klein (or Grassmann) representation [4] of lines in $\boldsymbol{CP}^3$. The *real* points of $\boldsymbol{CM}$ define the *compactified Minkowski space* M ($\supset M'$) and this is represented by the family of lines *in* $\mathcal{N}$. Thus, the light rays through a point O of M' correspond, in $\boldsymbol{CP}^3$, to the points of the line in $\mathcal{N}$ which represents O. Since this line has the holomorphic structure of $\boldsymbol{CP}^1$, we find agreement with the earlier discussion concerning the holomorphic structure of the light cone or sky of an observer at O.

Let $\boldsymbol{PT}^+$ and $\boldsymbol{PT}^-$ denote the portions of $\boldsymbol{CP}^3$ for which $\Sigma(Z^\alpha)$ is, respectively, positive or negative. Then the group of holomorphic self-transformations of $\boldsymbol{PT}^+$ is just the nonreflective conformal group $C_+^\uparrow(1,3)$ on M. This is related to the appropriate pseudo-unitary group on $\boldsymbol{C}^4$ (preserving $\Sigma(Z^\alpha)$) via the 4-1 local isomorphism

$$SU(2,2)\to C_+^\uparrow(1,3).$$

The Poincaré (or inhomogeneous Lorentz) group is a 10-real-parameter subgroup of $C_+^\uparrow(1,3)$ preserving I and a certain scaling for I.

The idea of the *twistor programme* [7] is to re-express the whole of basic physics in terms of the above space $\boldsymbol{CP}^3$ or the space $\boldsymbol{C}^4$ from which it arises. More correctly, however, twistor space should not be regarded as $\boldsymbol{C}^4$ (which is the co-ordinate space) but as a more directly physically defined space $\boldsymbol{T}$ whose elements (up to phase multipliers) correspond to *massless particles with spin*. The helicity is given by the value $\frac{1}{2}\hbar\Sigma(Z^\alpha)$. Moreover, it is envisaged that when the space-time is curved, or endowed with an electromagnetic field, or a more general gauge field, [8], the twistor space should acquire a more complicated complex structure than that of $\boldsymbol{C}^4$. Some considerable successes have been achieved, as it turns out, using this kind of method to describe such fields, in the cases for which the fields are *self-dual*. The general self-dual solution of Einstein's non-linear vacuum (Ricci flat) equations can, indeed, be completely coded [9] into the structure of a deformed version of a suitable portion of $\boldsymbol{T}$. Some recent ideas lead one to hope that it ought to be possible eventually to remove the self-dual restriction [11], but as yet this has not been satisfactorily achieved. The present situation with regard to the (non-linear) self-dual gauge fields (Yang–Mills fields) is broadly similar [12], [13], [24].

A remarkable feature of these constructions, when they *have* been successful, is that the information is not stored locally. The local structure of each of the deformed spaces is precisely the same as in the particular case when the field to be described

 The complex geometry of the natural world. *Proceedings of the International Congress of Mathematicians* (1978), pp. 189–94, *Acad. Sci. Fennica, Helsinki,* 1980

is zero! Nevertheless an infinite-dimensional system of inequivalent deformed spaces can be constructed and these correspond precisely to the freedom that one expects (free functions of three variables) for the fields in space-time. In the weak-field limit, we are concerned with infinitesimal deformations of complex structure. These are described in terms of the first holomorphic *sheaf cohomology groups* of portions of the flat twistor space T, so it appears that the twistor description of linear space-time fields should be described in terms of such cohomology groups. Indeed, for massless fields of any helicity this turns out to be the case [14], [10], [15]. The situation for massive fields is not so clear, but in any case products of twistor spaces are then needed, rather than the single space T. Part of the twistor programme is concerned with exploiting the extra freedom that then arises in order to obtain a classification of the numerous observed varieties of elementary particles [18], [19], [20].

One of the major aims of the twistor programme is to find an alternative version of quantum field theory which might, one hopes, avoid some of the difficulties of the conventional formulation. An aspect of this approach is that wave functions of massless fields, are described [11], [16], [17] as the elements of the sheaf cohomology groups $H^1(PT^+, \mathcal{O}(n))$ where $\mathcal{O}(n)$ refers to the sheaf of twisted holomorphic functions on PT^+, as defined by homogeneous functions on T of degree n. Here, $n=-2s\hbar^{-1}-2$, where s is the helicity of the field. There are many curious features of these descriptions which are only just beginning to be explored. For example it is known [21] that infinite-dimensional H^1's only arise here because the boundary $\mathcal{N}$ of PT^+ has a Levi form with one non-positive eigenvalue. It is also known [22], [23] that such infinite-dimensional H^1's then live *locally* on $\mathcal{N}$ (and, somewhat ironically, arise precisely because of the existence of *insoluble* differential equations of the Hans Lewy type). The space-time interpretation of these local H^1's is still somewhat obscure, but they appear to be concerned with the propagation of non-analytic behaviour along light rays. There is some hope also, that significantly different insights into the somewhat paradoxical nature of quantum-mechanical wave functions might eventually arise from the cohomological formalism that appears forced upon us in twistor descriptions.

References

1. R. Penrose, *The apparent shape of a relativistically moving sphere*, Proc. Cambridge Philos. Soc. **55** (1959), 137.

2. S. S. Chern and J. Moser, *Real hypersurfaces in complex manifolds*, Acta Math., **133** (1974), 219.

3. R. Penrose, *Twistor algebra*, J. Math. Phys. **8** (1967), 345.

4. W. V. D. Hodge and D. Pedoe, *Methods of algebraic geometry*. Vol. II. Cambridge Univ. Press, London, 1951.

5. R. Penrose and M. A. H. MacCallum, Twistor theory, an approach to the quantization of fields and space-time, Phys. Rep. **6C** (1973), 241.

6. R. Penrose, *Twistor theory: its aims and achievements*, Quantum Gravity an Oxford Symposium, C. J. Isham, R. Penrose and D. W. Sciama, eds., Oxford Univ. Press, London, 1975.

7. *The twistor programme*, Rep. Mathematical Phys. **12** (1977), 65.

8. J. C. Taylor, *Gauge theories of weak interactions*, Cambridge Univ. Press, London, 1975.

9. R. Penrose, *Non-linear gravitons and curved twistor space*, General Relativity and Grav. **7** (1976), 31.

10. *Twistor quantization and curved space-time*, Internat. J. Theor. Phys. **1** (1968), 61.

11. *On the twistor description of massless fields*, Complex Manifold Techniques in Theoretical, Physics, D. Lerner and P. Sommers, eds. Pitman Publ. Co., London, 1979.

12. R. S. Ward, *On self-dual gauge fields* Phys. Lett. **61A** (1977), 81.

13. M. F. Atiyah and R. S. Ward, *Instantons and algebraic geometry*, Comm. Math. Phys. (1977)

14. R. Penrose, *Solutions of the zero rest-mass equations*, J. Math. Phys. **10** (1969), 61.

15. Twistor Newsletter **2** (1976); **5** (1977).

16. R. O. Wells, Jr., *Complex manifolds and mathematical physics*, Bull. Amer. Math. Soc. **1** (1979), 296—336.

17. S. Gindikin and G. Henkin, *Transformation de Radon pour la d″-cohomologie des domaines q-linéairement concaves*, C. R. Acad. Sci. Paris Ser. A. **303** (1978).

18. R. Penrose, *Twistors and particles*, Quantum Theory and the Structures of Time and Space, L. Castell, M. Drieschner, C. F. von Weiszacker, eds., Hanser, Munich, 1975.

19. Z. Perjés, *Twistors and particles*, Phys. Rev. D **11** (1975), 2031.

20. *Perspectives of the Penrose theory of particles*, Reports on Math. Phys. **12** (1977).

21. A. Andreotti and H. Grauert, *Théorèmes de finitude pour la cohomologie des espaces complexes*, Bull. Soc. Math. France **90** (1962), 193.

22. A. Andreotti and C. D. Hill, *E. E. Levi convexity and the Hans Lewy problem*, Ann. Scuola. Norm. Sup. Pisa CI. Sci. **26** (1972), 325; 747.

23. C. D. Hill, *Partial differential equations*, Proc. Sympos. Pure Math., vol. 23, Amer. Math. Soc., Providence, R. I., 1971, p. 135.

24. M. F. Atiyah, N. J. Hitchin, V. G. Drinfeld and Yu. I. Manin, *Construction of instantons*, Phys. Lett. **65A** (1978), 185.

116
ON THE EVALUATION OF TWISTOR COHOMOLOGY CLASSES

Twistor Newsletter **10** (1980). Also published in *Further Advances in Twistor Theory*, Vol. 1. Eds L. J. Mason and L. P. Hughston, Longman, 1990

 On the evaluation of twistor cohomology classes. *Twistor Newsletter* **10** (1980). Also published in *Further Advances in Twistor Theory*, Vol. 1. Eds L. J. Mason and L. P. Hughston, Longman, 1990

§I.5.4 On the evaluation of twistor cohomology classes *by R.Penrose* (TN 10, July 1980)

In the programme for evaluating twistor diagrams initiated by Ginsberg & Huggett much use is made of the *evaluation map*

$$H^n(M,\Omega^n)\to\mathbb{C} \tag{a}$$

where M is a compact n-dimensional complex manifold and Ω^n the sheaf of holomorphic n-forms on M. (This is the map used in Serre duality, but is, by itself, a more elementary thing). The idea is to piece together various cohomology group elements correpsonding to different portions of the diagram, making much use of the 'multiply map' operation

$$\bullet: H^p(\mathcal{U},\mathcal{F})\times H^q(\mathcal{V},\mathcal{T})\to H^{p+q+1}(\mathcal{U}\cup\mathcal{V},\mathcal{F}\times\mathcal{T}), \tag{b}$$

the aim being, in each case, to fit together the entire compact manifold (generally $\mathbb{CP}^3\times\mathbb{CP}^3\times\cdots\times\mathbb{CP}^3$) out of $\mathcal{U}$, $\mathcal{V}$, $\cdots$ and finally to make use of (a) to evaluate the completed diagram. For the simple (scalar product) diagram $\boxed{f}\!\!\bullet\!\!\triangleleft\boxed{g}$, where $f\in H^1(\overline{\mathbb{PT}^+},\mathcal{O}(r-2))$, $g\in H^1(\overline{\mathbb{PT}^-},\mathcal{O}(-r-2))$, Eastwood has suggested the alternative procedure of forming the *cup* product

$$f\cup g\in H^2(\mathbb{PN},\Omega^3) \tag{c}$$

where Ginsberg & Huggett would have

$$f\bullet g\in H^3(\mathbb{PT},\Omega^3). \tag{d}$$

The evaluation of (c) to obtain the complex number equal to the amplitude $\boxed{f}\!\!\bullet\!\!\triangleleft\boxed{g}$ can be achieved either by Dolbeault means (which, however, does not make its essentially holomorphic character obvious (to me) or by a Čech (or abstract) procedure. In fact, the evaluation of (c) and (d) yield the same (correct) answer.

This alternative method suggests the use of an obvious generalization

$$H^p(M,\Omega^n)\xrightarrow{V}\mathbb{C} \tag{e}$$

where M is an n-dimensional complex manifold, now *not* necessarily compact, and where V is a *compact* $(n+p)$-real-dimensional submanifold of M (or, more generally, an $(n+p)$-cycle in M) $(0\le p\le n)$. Again this has an obvious Dolbeault interpretation (not immediately obviously holomorphic) or a Čech (or abstract nonsense) interpretation obtained by using

$0\to\mathbb{C}\to\Omega^0\to\Omega^1\to\cdots\to\Omega^n\to 0$ to derive a sequence (non-exact) of maps

$$H^p(M,\Omega^n)=H^p(M,d\Omega^{n-1})\to H^{p+1}(M,d\Omega^{n-2})\to\cdots \qquad (f)$$

$$\cdots\to H^{p+n-1}(M,d\Omega^0)\to H^{p+n}(M,\mathbb{C})\xrightarrow{V}\mathbb{C}$$

the last map being ordinary (Čech) cohomology evaluation on the cycle V. This incorporates two special cases: (i) good old-fashioned contour integration (given when $p=0$ and $V=$ the contour); (ii) the original evaluation map (a) (given when $p=n$ and $V=M$ (compact)).

One powerful advantage of this is that it enables us to treat cases when M is non-compact. (Previously we had had to resort to using compact cohomology for (a) when M is non-compact, but this is not so transparent.) The disadvantage, of course, is that we need to specify V. Two examples of the use of (e) spring to mind: (I) non-projective twistor integrals, e.g. the

$$\frac{1}{(2\pi i)^{2n}}\oint\cdots f(\underset{i}{Z})\cdots d^2\underset{1}{\pi}\wedge\cdots\wedge d^2\underset{n}{\pi}$$

that fit in with the twistor particle programme. (I have checked this explicitly for the normal 1-twistor function evaluation at a field point; here $V\equiv S^3$). (II) Evaluation of a twistor function (H^1) at a googly field point—which amounts effectively to $\boxed{f}\!\!\bullet\!\!\overset{p}{\underset{q}{<}}\!\!\boxed{x}$, where $=\!\!\overset{p}{\underset{q}{}}\!\!\boxed{x}$ is the cohomology class (H^1) which arises from a googly map $x:\mathcal{T}\to\mathbf{S}$ (ω-spin space), $=\!\!\overset{p}{\underset{q}{}}\!\!\boxed{x}$ being

$$x^{-1}\Big(\frac{d^2\omega}{(\omega^0)^{p+1}(\omega^1)^{q+1}}\Big).$$

We have $=\!\!\overset{p}{\underset{q}{}}\!\!\boxed{x}=\overset{p}{\underset{q}{}}\frac{A}{B}$ in ordinary twistor diagrams, the line x in $\mathbb{PT}$ being the intersection of planes A and B. Here $\mathcal{T}$ is asymptotic twistor space, which need not be compact. Deforming slightly away from the flat case, we find that a suitable V exists[1] with $V\simeq S^3\times S^3$.

Finally, it is not hard to see that any evaluation using a multiply map can also be achieved using a cup product, the specific V for '$\cup$' arising as the intersection of the V for '$\bullet$' with $\partial\mathfrak{U}$, or, homologously, with $\partial\mathcal{V}$; '$\cup$' is more flexible, though '$\bullet$' is more unique.

[1]**Note added, July 1989:** The situation turns out to be considerably more complicated than is suggested here. The $S^3\times S^3$ contour does not actually survive when the googly maps are deformed from those of flat space to those arising in specific self-dual curved models. It appears that contours of a Pochhammer type are likely to be needed.

References

Ginsberg, M.L. & Huggett, S.A., §O.5.12 in *Advances in twistor theory*

Penrose, R. (1975) in *Quantum gravity; an Oxford symposium*, eds. C.J.Isham, R.Penrose and D.W.Sciama O.U.P.

Penrose, R. §O.3.9.

Postcript: We have the following obvious (!) generalization of (e):

$$H^p(M, d\Omega^{q-1}) \xrightarrow{V} \mathbb{C}, \tag{g}$$

where M is complex n-dimensional and V is real $(p+q)$-dimensional. ($d\Omega^{q-1}$ = sheaf of holomorphic *closed* q-forms.) Case $q=n$ gives (e). **Proof**, the same. '***Application***': ordinary twistor integral $\frac{1}{2\pi i}\oint \cdots f(Z)\pi_{A'} d\pi^{A'}$ viewed as $S^1(=V)$ in $\mathbb{C}^2$, $p=0$, $q=1$, $n=2$.

117
THREE CHANNELS FOR THE BOX DIAGRAM

Twistor Newsletter **10** (1980). With S. A. Huggett. Also published in *Further Advances in Twistor Theory*, Vol. 1. Eds L. J. Mason and L. P. Hughston, Longman, 1990

 Three Channels for the Box Diagram. *Twistor Newsletter* **10** (1980). Also published in *Further Advances in Twistor Theory*, Vol. 1. Eds L. J. Mason and L. P. Hughston, Longman, 1990

§1.5.5 Three channels for the box diagram *by S.A.Huggett & R.Penrose* (TN 10, July 1980)

Introduction. In this article we show that all three channels for a ϕ^4 integral do in fact survive translation into twistors. (We do not use any cohomological techniques, and nor do we blow the box diagram up.) This result refutes the ancient belief that the box diagram only has one of these channels and vindicates the hope that twistor diagrams incorporate crossing symmetry [1].

Translation Procedure. The massless scalar ϕ^4 integral is $\int_S \phi_1\phi_2\phi_3\phi_4 d^4x$ (1)

where S is a 4 real dimensional contour in $\mathbb{CM}$. Suppose the zero rest mass fields $\phi_1,\cdots,\phi_4$ are elementary states. The first step in the translation is to choose the following twistor functions to generate these elementary states

$$\frac{1}{\begin{matrix}W & W\\ | & |\\ A & B\end{matrix}} \leftrightarrow \phi_1,\quad \frac{1}{\begin{matrix}C & D\\ | & |\\ X & X\end{matrix}} \leftrightarrow \phi_2,\quad \frac{1}{\begin{matrix}Y & Y\\ | & |\\ E & F\end{matrix}} \leftrightarrow \phi_3,\quad \frac{1}{\begin{matrix}G & H\\ | & |\\ Z & Z\end{matrix}} \leftrightarrow \phi_4$$

Given $A,B,\cdots,H$ in a particular position move them until one of the following conditions holds:

channel $<12|34>$: ϕ_1,ϕ_2 negative frequency and ϕ_3,ϕ_4 positive
channel $<13|24>$: ϕ_1,ϕ_3 negative frequency and ϕ_2,ϕ_4 positive
channel $<14|23>$: ϕ_1,ϕ_4 negative frequency and ϕ_2,ϕ_3 positive

Now (i) draw in the contour $S=\mathbb{M}$ and (ii) continuously move $A,B,\cdots,H$ *and* the contour until $A,B,\cdots,H$ are in their original positions. We now have three space-time contours, one for each channel. We use the notation $W_\alpha=(W_A,W^{A'})$ so that (1) becomes

$$\int \frac{W^A dW_A \wedge X^{A'} dX_{A'} \wedge Y^B dY_B \wedge Z^{B'} dZ_{B'} \wedge d^4x}{\rho_x \begin{pmatrix} W & W & C & D & Y & Y & G & H \\ | & | & | & | & | & | & | & | \\ A & B & X & X & E & F & Z & Z \end{pmatrix}} \tag{2}$$

where ρ_x restricts all the twistors to go through x^a. This integral is over a contour in the 8 complex dimensional space $\mathbb{B}$. $\mathbb{B}$ is a bundle over $\mathbb{CM}$ with fibre

$$\{(W_A,X_{A'},Y_A,Z_{A'})\in(\mathbb{CP}^1)^4\}.$$

We define two subbundles of $\mathbb{B}$.

$$\mathbb{I}=:\{(W_A,X_{A'},Y_A,Z_{A'},x^a)\in\mathbb{B}: W_A\propto Y_A \text{ and } X_{A'}\propto Z_{A'}\}$$

$$\mathbb{U}=:\{(W_A,X_{A'},Y_A,Z_{A'},x^a)\in\mathbb{B}: W_A\propto Y_A \text{ or } X_{A'}\propto Z_{A'}\}$$

The next step in the translation is the projection p from $\mathbb{B}$ to

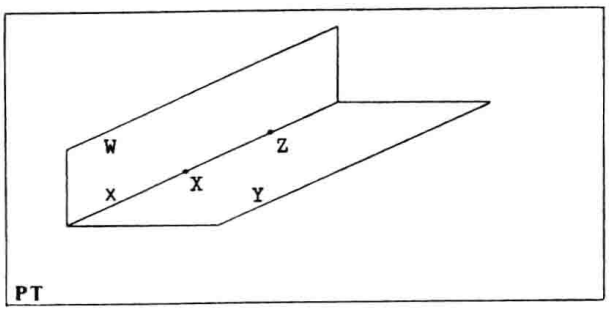
$$\left\{ \begin{matrix} | \\ W \\ | \end{matrix} ; \begin{matrix} | \\ X \\ | \end{matrix} ; \begin{matrix} | \\ Y \\ | \end{matrix} ; \begin{matrix} | \\ Z \\ | \end{matrix} : \begin{matrix} W \\ | \\ X \end{matrix} = \begin{matrix} W \\ | \\ Z \end{matrix} = \begin{matrix} Y \\ | \\ X \end{matrix} = \begin{matrix} Y \\ | \\ Z \end{matrix} = 0 \right\} =: \mathcal{S} \tag{*}$$

given by

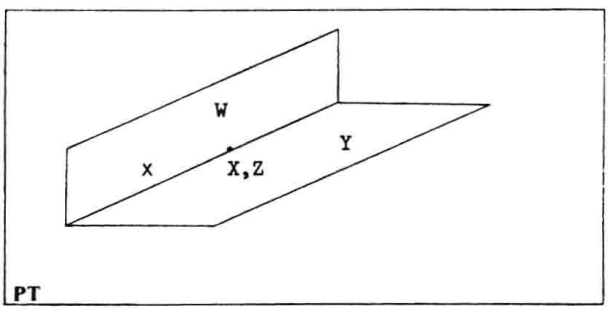
$$p(W_A,\ X_{A'},Y_A,Z_{A'},x^a)=(W_A,-ix^{AA'}W_A;ix^{AA'}X_{A'},X_{A'};\ Y_A,-ix^{AA'}Y_A;\ ix^{AA'}Z_{A'},Z_{A'}).$$

 Three Channels for the Box Diagram. *Twistor Newsletter* **10** (1980). Also published in *Further Advances in Twistor Theory*, Vol. 1. Eds L. J. Mason and L. P. Hughston, Longman, 1990

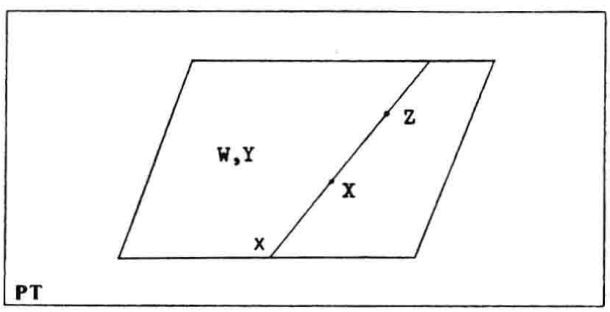
$p(\mathbb{B}-\mathbb{U})$



$p(\mathbb{U}-\mathbb{I})$ looks like this



or this



This projection is a biholomorphism on $\mathbb{B}-\mathbb{I}$, but consider

$$p(\mathbb{I}) = \{ W;X;Y;Z : W \propto Y, X \propto Z, \text{ and } \begin{matrix} W \\ X \end{matrix} = 0 \}$$

which (as was realised after a question from M.G.Eastwood) is singular in $\mathcal{PT}$. $\mathbb{I}$ has 6 complex dimensions and is a blowup of $p(\mathbb{I})$ which only has 5. Therefore $\mathbb{B}$ is a blow up of $\mathcal{PT}$. (Actually, for $\mathbb{B}$ to be a standard blow up of $\mathcal{PT}$ along $p(\mathbb{I})$ the inverse image of a point in $p(\mathbb{I})$ would have to be a $\mathbb{CP}^2$, because $\operatorname{codim}_{\mathbb{C}} p(\mathbb{I}) = 3$. In other words the space of normal directions to $p(\mathbb{I})$ in $\mathcal{PT}$ is $\mathbb{CP}^2$. In fact not *all* these directions are separated in the inverse image, which is only a $\mathbb{CP}^1$. So $\mathbb{B}$ is a more general *resolution* of $\mathcal{PT}$. Thanks to N.J.Hitchin for discussion on this point.)

The differential form in (2) is nonzero on $\mathbb{I}$ so when we blow $\mathbb{I}$ down this form will be singular on $p(\mathbb{I})$. Therefore if a contour in $\mathbb{B}$ is to survive the translation it must avoid $\mathbb{I}$, so that the restriction of the form to this contour remains nonsingular. The last step in

obtaining the box diagram is to go from the space $\mathcal{PT}-p(\mathbb{I})$ to the space

$$\{W;X;Y;Z\colon \begin{matrix} | \\ W \\ | \\ X \end{matrix} \neq 0, \begin{matrix} | \\ W \\ | \\ Z \end{matrix} \neq 0, \begin{matrix} | \\ Y \\ | \\ X \end{matrix} \neq 0, \begin{matrix} | \\ Y \\ | \\ Z \end{matrix} \neq 0\} =: \mathcal{B}$$

using Cauchy's integral formula once for each subspace of the form $\left\{\begin{matrix} | \\ W \\ | \\ X \end{matrix} = 0\right\}$. Suppose γ is a contour in $\mathcal{PT}-p(\mathbb{I})$. Then this last stage maps γ to the contour $\delta^4(\gamma)$ in $\mathcal{B}$, where δ is Leray's cobord map. We construct $\delta^4(\gamma)$ in the last section.

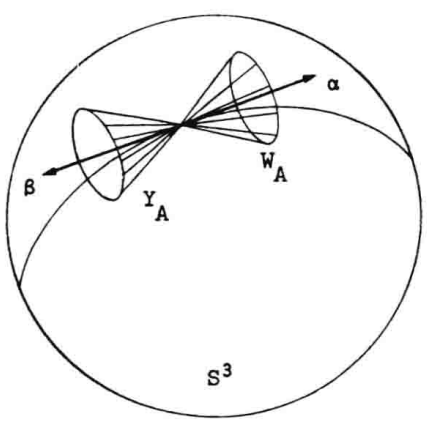
Proof that the three contours avoid $\mathbb{I}$

Easy case—channel $<13|24>$: We can start with a particular location for $ABC\cdots H$ and then move them away from that location (keeping the lines AB and EF in $\mathbb{PT}^+$, for negative frequency, and the lines $C\cap D$ and $G\cap H$ in $\mathbb{PT}^-$, for positive frequency). Think of $\mathbb{M}$ as the Einstein universe identified: $S^3\times S^1$. Choose $A=E$, $B=F\in\mathbb{PT}^+$ in such a way that the Robinson congruence α defined by A points (when projected spatially) along a right rotation of S^3 and that β defined by B along the opposite rotation of S^3.

Similarly, choose $\overline{C}=\overline{G}$, $\overline{D}=\overline{H}\in\mathbb{PT}^+$ so that C and D define γ and δ whose spatial projections are also opposite right rotations on S^3. We need a contour for the space

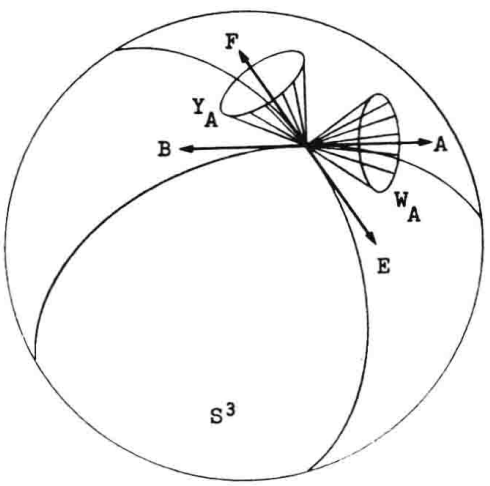
$$p^{-1}\left(\mathcal{PT}\cap\left\{\begin{matrix} W & W & C & D & Y & Y & G & H \\ | & | & | & | & | & | & | & | \\ A & B & X & X & E & F & Z & Z \end{matrix} \neq 0\right\}\right)$$

avoiding $\mathbb{I}$. Actually, in this case we can even avoid $\mathbb{U}$. For each x^a we need $(W_A,X_{A'},Y_A,Z_{A'})$ such that the flagpoles of $W_A,X_{A'},Y_A$, and $Z_{A'}$ avoid the Robinson congruences α,β,γ, and δ respectively, and to miss $\mathbb{U}$ we also need $W_A\not\propto Y_A$ and $X_{A'}\not\propto Z_{A'}$. Easy. Let the W_A flagpole (when projected into S^3) execute a narrow cone about the α-direction and let the Y_A flagpole do the same (independently) about the β-direction (see picture).



We get $S^1\times S^1$'s worth of contour at each point of the S^3. Treat $X_{A'}$ and $Z_{A'}$ similarly (and independently). Since these Robinson congruences are constant in Einstein (S^1) time we have $S^3\times S^1$'s worth of $S^1\times S^1\times S^1\times S^1$ contours avoiding $\mathbb{U}$. This is the required 8 real dimensional contour.

Harder cases—channels $<12|34>$ *and* $<14|23>$: For channel $<12|34>$, $A,B,\overline{G}$ and $\overline{H}$ are in $\mathbb{PT}^+$ and $E,F,\overline{C}$ and $\overline{D}$ are in $\mathbb{PT}^-$. Now we cannot choose $A=E$ because they correspond to a right rotation and a left rotation respectively. So although we can make W_A and Y_A execute narrow cones about the directions on S^3 corresponding to A and F respectively (as before, remember F *was* equal to B) Y_A and W_A will encounter one another when these directions get close. This will happen near a great circle on the S^3. Similarly for $X_{A'}$ and $Z_{A'}$—they will encounter one another near another great circle on the S^3. So all we have to do is make sure that these great circles do not intersect. Then at no stage do we have $W_A\propto Y_A$ *and* $X_{A'}\propto Z_{A'}$, so that $\mathbb{I}$ is avoided for all x^a on the S^3, as required. Channel $<14|23>$ is similar, but notice that the channels $<12|34>$ and $<14|23>$ *do* intersect $\mathbb{U}$.



Getting from $\mathcal{PT}$ *to* $\mathcal{B}$. Suppose γ is a contour in $\mathcal{PT}-p(\mathbb{I})$ and suppose for now that it also avoids $p(\mathbb{U})$ (as does the $<13|24>$ contour). A point on γ is given by two points and two planes as drawn in the picture of $p(\mathbb{B}-\mathbb{U})$. Fix the planes W and Y and consider the neighbourhood of X in any plane through X not containing Y. The planes W and Y never coincide on γ so we can take X off them and impose the conditions $\begin{matrix}W\\|\\X\end{matrix}=\begin{matrix}Y\\|\\X\end{matrix}=0$ using two

Cauchy integrals. Similarly for the point Z we use two more Cauchy integrals, so that we integrate over four S^1's in all before using γ. Now suppose γ intersects $p(\mathbb{U})$. The above procedure still works when $X \propto Z$, but if $W \propto Y$ we have to dualise (draw the whole thing in dual twistor space and fix the planes X and Z, and so on). In the previous section we saw that γ never intersects $p(\mathbb{I})$, so we can always use one of these procedures.

References

[1] Hodges, A.P. (1979) 'Crossing and twistor diagrams' in §0.5.6.

[2] Huggett, S.A. (1980) D.Phil. thesis; Oxford.

[3] Huggett, S.A. (1979) 'Blowing up the box' in §0.5.10.

Additional note: The cohomological status of the two 'hard' contours is still mysterious. See Huggett, S.A. and Singer, M.A. 'relative cohomology and projective twistor diagrams' (to appear in Trans. A.M.S.) for a discussion of this point.

118
PUZZLE: A SEQUENCE FOR TWISTORIANS

Twistor Newsletter **10** (1980)

For the solution, see Chapter 251.

A Sequence for Twistorians

Fill in the missing number:

..., 7, 9, 12, (?), 24, 36, 56, 90, ...

Roger Penrose

119
A BRIEF INTRODUCTION TO TWISTORS

Surveys in high energy physics **1** (1980), no. 4, 267–88

Surveys in High Energy Physics
Volume 1(4), August 1980, pp. 267-288
0142-2413/80/0104 0267$04.50/0
©1980 harwood academic publishers gmbh
Printed in the United States of America

A BRIEF INTRODUCTION TO TWISTORS

R. PENROSE
Mathematical Institute
24-29 St. Giles
Oxford

Abstract

Twistor theory provides a formalism for physics different from (though initially equivalent to) the standard space-time description. Space-time points are regarded as derived, non-local, objects in twistor space, which (in the initial theory) is a complex 4-dimensional vector space. A twistor describes a (classical) spinning massless particle, while holomorphic functions of twistors encode the space-time field equations for massless fields of each different spin.

§1. Motivations

For about sixteen years, the theory of twistors has gradually evolved as an attempt at a reformulation of basic physics. The theory is still very far from being a coherent or very comprehensive description of Nature. But it does, even in its present incomplete state, offer a number of intriguing and, in some ways, remarkable descriptions of various different but interconnected parts of physics. Some of the main developments in these areas are described in the accompanying articles[1,2,3,4]. My purpose here is to provide some motivation for the theory and to

outline its basic geometrical and physical ideas.

There are several different ingredients to the motivations behind the twistor approach. Perhaps prime among these is the feeling that the concept of space-time point, so basic to the conventional approach to physics, is an abstraction rather far removed from the realities of the natural world. Even at the level of classical physics, a "point particle" is not described by a space-time point but by a one-dimensional space-time curve. But more significant is the fact that at the quantum level, our descriptions of phenomena in space-time terms become, at best, somewhat blurred. Owing to the Heisenberg uncertainty principle and quantum linear superposition, our classical pictures of physical objects moving smoothly about from one definite position to another have now to be abandoned as accurate representations of reality. Such pictures retain a heuristic value only, and we proceed from them using certain formal rules that go under the general heading of "quantization". Once these rules have been incorporated into our modes of thought, there is the tendency to feel that the concept a classical background space-time can be retained, the procedure of "quantization" being applied only to its contents. We do know from Einstein's general relativity that this cannot ultimately be the case, since the detailed structure of space-time is dependent upon the nature of its contents. But the standard view on this question has been that only at the Planck-Wheeler level of 10^{-33}cm would a quantum nature for the structure of space-time begin to show up. However, this presupposes that a good physical concept of space-time has reality right down to that level. It is my own view that at those much larger levels at which quantum

phenomena already provide us with difficulties in picturing reality, the concept of classical space-time has begun to become inappropriate. While it may be that only at 10^{-33}cm are we forced to abandon the picture of classical space-time it may nevertheless be the case that the use of conventional space-time pictures is a hindrance to progress even at the level of, say, hadron physics, or larger.

But it would, of course, be folly simply to abandon the highly successful descriptions of conventional physics. The primary concern of any aspiring new theory must be that it reproduces, in some form, all the successes of existing theory. So any new geometry that is invoked must be, in some strong sense, equivalent to that of standard physics. Yet it must also contain the potentiality to provide us with something qualitatively completely new, if the new geometrical framework is to have any profound advantages over the old.

Since it is quantum physics that has led us to seek a new geometry for the world, it is in the formalism of quantum mechanics that we must seek clues as to the nature of this new geometry. Now, the primary mathematical ingredient of quantum mechanics is the complex number field. For it is complex numbers that form the coefficients in quantum linear superposition, or, equivalently, the probability amplitudes or matrix elements that are so fundamental to the theory. (And while one can, of course, write all equations out in their real and imaginary parts, the fact remains that the "quantum spread" of two states is always two-real-dimensional, never one-real-dimensional, nor three, nor more.) And "i" plays a crucial role in the Schrödinger equation, etc.

These complex numbers also interrelate with the

structure of space-time. This is most easily seen for the quantum states of a spin ½ massive particle. The sphere of directions, in which a spin-axis can point, corresponds precisely to the array of pure states for such a particle, and provides a realization of the Riemann sphere representing the ratios of pairs of complex numbers (probability amplitudes). This is reflected in the local isomorphism between the spatial rotation group SO(3) and the quantum group SU(2). Since all quantum superpositions are connected with one another via interference phenomena, this entails a universal interconnection, however tenuous it may seem at first, between quantum linear superposition and the structure of space - and, notably, with the three-dimensionality of space, because of the physical significance of the above local isomorphism $SU(\overset{2}{3}) \to SO(\overset{3}{2})$.

But, of course, we must consider relativity, and the relativistic version of the above local isomorphism, which is $SL(2,\mathbb{C}) \to O^{\uparrow}_{+}(1,3)$. The connection with quantum superposition is less clear than before, but again the Riemann sphere plays a role, this time as the space of light rays through a point in space-time, i.e. as the celestial sphere. If we think of this space as a complex one-dimensional manifold, then we recognize the restricted Lorentz group $O^{\uparrow}_{+}(1,3)$ as its group of holomorphic (i.e. complex-analytic) self-transformations.

We are now in a position to consider, as a first physical approximation to twistor theory[5], that instead of space-time points being regarded as the primary physical elements, entire light rays (null straight lines in Minkowski space $\mathbb{M}^{I}$) should be so regarded. We represent each light ray as a point in a new space $\mathbb{PN}$. Those light

rays that pass through a fixed point in $\mathbb{M}^I$ represent points which constitute a complex 1-space (Riemann sphere), so we may ask whether the entire space $\mathbb{PN}$ can, in some way, be regarded as a higher-dimensional complex manifold. In fact this cannot be so, since $\mathbb{PN}$ is 5-dimensional while complex manifolds, when interpreted as real spaces, must be even-dimensional. But, remarkably, it turns out that $\mathbb{PN}$ is imbedded as a real hypersurface in a certain 6-real-dimensional manifold $\mathbb{PT}$ - referred to as projective twistor space - which is, in a natural way, a complex 3-manifold. In fact, $\mathbb{PT}$ is a complex projective 3-space, and the complex 1-spaces that represent the light cones of

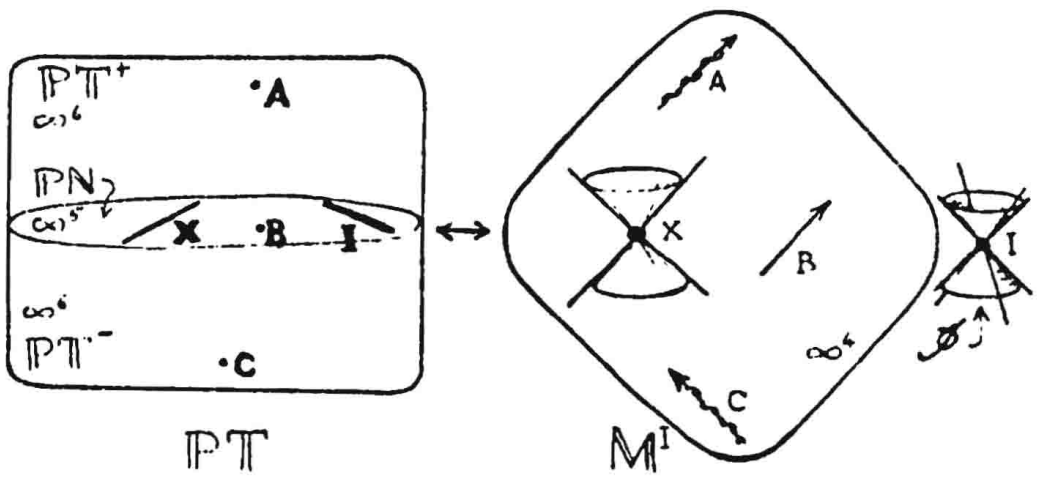


Figure 1:
Projective twistor space $\mathbb{PT}$ and its relation to Minkowski space $\mathbb{M}^I$.

of $\mathbb{M}^I$ are complex projective lines lying in $\mathbb{PN}$. To make the correspondence more complete, we should adjoin to $\mathbb{M}^I$ a light cone $\mathscr{I}$ at infinity, together with its vertex I, which makes $\mathbb{M}^I$ into a compact manifold $\mathbb{M}$. The cone $\mathscr{I}$ (or, equivalently, I) is represented by a particular projective line $\mathbf{I}$, lying in $\mathbb{PN}$ (fig. 1).

As a second physical approximation[6], it is possible to provide a physical interpretation for the points of $\mathbb{PT}$ not lying on $\mathbb{PN}$. In fact $\mathbb{PT}$ is divided into two regions, labelled $\mathbb{PT}^+$ and $\mathbb{PT}^-$ by the real hypersurface $\mathbb{PN}$. In a well-defined sense, to be described in §2, each point of $\mathbb{PT}^+$ can be identified with a right-handed classical spinning photon, and each point of $\mathbb{PT}^-$ with a left-handed one. The points of $\mathbb{PN}$ now emerge as describing high-energy limits of such particles (except for the points of $\mathbb{X}$, which may be thought of as low-energy limits). Only when the energy is high, can the photons be adequately localized as single light rays.

But these descriptions refer only to projective twistors. A twistor proper is an element of a 4-dimensional complex vector space $\mathbb{T}$, the 1-dimensional linear subspaces of which are the points of the projective space $\mathbb{PT}$. As a third physical approximation[7], we find that a twistor Z^α can be thought of as a sort of "Hermitian square-root" of the kinematical structure (energy-momentum-angular-momentum structure) of a classical massless particle of arbitrary (real) helicity - being determined up to the phase freedom $Z^\alpha \to e^{i\theta} Z^\alpha$ (θ real) by this structure. The points of $\mathbb{N} \subset \mathbb{T}$ (from which $\mathbb{PN}$ arises) describe zero helicity particles.

A twistor is, however, more correctly a quantum object rather than a classical one. As a fourth physical approximation[8], we may, in effect, interpret a twistor as defining a certain type of linear operator that acts upon the wave-function of a massless particle of helicity s and produces a new state of helicity s + ½ħ. (It can also act on massive particle states and many-particle states.) A dual twistor (or complex conjugate of a twistor) similarly

defines a helicity lowering operator, of a precisely analogous kind. With this interpretation, the correct (canonical) commutation rules arise, between a twistor and its complex conjugate.

All these descriptions refer only to univalent (or single-indexed) twistors. Twistors of higher valence are constructed from tensor products of these. Furthermore, it appears that when interactions are present, the twistor space becomes deformed and loses its linear structure (the projective twistor space sometimes even becoming a non-Hausdorff complex manifold). Its holomorphic (complex-analytic) structure persists, however, and it seems that pure "holomorphicity" can, in a strange non-local way, take over the role played by field equations in the normal space-time descriptions, at least in a surprising number of instances[1,2,10,11,12]. One aim of the twistor programme, in fact, is to extend the scope of twistor holomorphicity so as to incorporate as many of the field equations of physics as possible.

As twistor theory stands at present, we have found ourselves led into certain rather abstruse areas of mathematics (complex algebraic geometry, sheaf cohomology, deformation theory and several complex variable analysis), while many different-looking mathematical descriptions of a twistor have arisen (Robinson congruences, α-planes, complex null geodesics at infinity, elements of the basic representation space of SU(2,2)). It is to be hoped that with further understanding, the physical basis of the theory will become clearer. So far, the directions of development in the theory have been much guided by the mathematics. It may be that this has carried the theory some way from its original aims.

Nevertheless, the mathematics often seems compelling and it has led to some striking and unexpected interrelations between complex geometry and physical laws.

§2. Explicit Description of a Twistor

According to our "third physical approximation", the components $(z^0,z^1,z^2,z^3) \in \mathbb{C}^4$ of twistor $z^\alpha \in \mathbb{T}$, together with the components

$$(\bar{z}_0,\bar{z}_1,\bar{z}_2,\bar{z}_3) = (\overline{z^2},\overline{z^3},\overline{z^0},\overline{z^1}) \qquad 2.1$$

of its complex conjugate dual twistor $\bar{z}_\alpha$, determine the components of the momentum p^a and angular momentum M^{ab} (with respect to an origin O in Minkowski space $\mathbb{M}^I$) of a massless free particle, according to

$$\begin{aligned} p^0 &= \tfrac{1}{\sqrt{2}}(z^3\bar{z}_1 + z^2\bar{z}_0),\ p^1 = \tfrac{1}{\sqrt{2}}(z^3\bar{z}_1 - z^2\bar{z}_0) \\ p^2 &= \tfrac{1}{\sqrt{2}}(-z^2\bar{z}_1 - z^3\bar{z}_0),\ p^3 = \tfrac{i}{\sqrt{2}}(z^2\bar{z}_1 - z^3\bar{z}_0) \end{aligned} \qquad 2.2$$

$$\begin{aligned} M^{01} &= -M^{10} = \tfrac{i}{2}(z^0\bar{z}_0 - z^1\bar{z}_1 - z^2\bar{z}_2 + z^3\bar{z}_3) \\ M^{02} &= -M^{20} = \tfrac{i}{2}(z^0\bar{z}_1 + z^1\bar{z}_0 - z^2\bar{z}_3 - z^3\bar{z}_2) \\ M^{03} &= -M^{30} = \tfrac{1}{2}(z^0\bar{z}_1 - z^1\bar{z}_0 - z^2\bar{z}_3 + z^3\bar{z}_2) \\ M^{12} &= -M^{21} = \tfrac{i}{2}(z^0\bar{z}_1 - z^1\bar{z}_0 + z^2\bar{z}_3 - z^3\bar{z}_2) \\ M^{32} &= -M^{23} = \tfrac{1}{2}(z^0\bar{z}_0 - z^1\bar{z}_1 + z^2\bar{z}_2 - z^3\bar{z}_3) \\ M^{13} &= -M^{31} = \tfrac{1}{2}(z^0\bar{z}_1 + z^1\bar{z}_0 + z^2\bar{z}_3 + z^3\bar{z}_2) \\ M^{00} &= M^{11} = M^{22} = M^{33} = 0. \end{aligned} \qquad 2.3$$

In addition to the required properties $p_a p^a = 0$, $p^0 > 0$ automatically holding, the Pauli-Lubanski spin vector

$$S_a = \tfrac{1}{4} e_{abcd} p^b M^{cd}$$

also automatically satisfies the necessary relation

$$S_a = s p_a ,$$

where the helicity s is given by

$$s = \tfrac{1}{2}(z^\alpha \bar{z}_\alpha + \bar{z}_\alpha z^\alpha) . \qquad 2.4$$

In (2.4) (and also (2.2), (2.3)), the non-commutativity that is needed in the case of quantum operators is being allowed for. The required commutators are

$$[z^\alpha, z^\beta] = 0, \ [\bar{z}_\alpha, \bar{z}_\beta] = 0 \text{ and } [z^\alpha, \bar{z}_\beta] = \hbar \delta^\alpha_\beta . \qquad 2.5$$

Substituting (2.5) into (2.2) and (2.3), we get the standard commutation relations for the generators of the Poincaré group.

The (classical) twistor z^α is determined up to the phase transformation $z^\alpha \to e^{i\theta} z^\alpha$ (θ real) by the p^a, M^{ab} of a (classical) massless particle and, conversely, every such twistor (except those for which $z^2 = z^3 = 0$) is related to a (finite) massless particle in this way. Projective twistor space $\mathbb{P}\mathbb{T}$ is the space of complex proportionality classes of non-zero (classical) twistors. The standard coordinates for $\mathbb{P}\mathbb{T}$ are the three independent complex ratios $z^0 : z^1 : z^2 : z^3$. $\mathbb{P}\mathbb{T}'$ is divided into two complex 3-manifolds $\mathbb{P}\mathbb{T}^+$ (given by $z^\alpha \bar{z}_\alpha > 0$), which describes right-handed massless particles, and $\mathbb{P}\mathbb{T}'^-$ (given by $z^\alpha \bar{z}_\alpha < 0$) describing left-handed ones, by their

common 5-real-dimensional boundary $\mathbb{PN}$ (given by $Z^{\alpha}\bar{Z}_{\alpha} = 0$). Similarly, twistor space $\mathbb{T}$ is divided into $\mathbb{T}^{+}$ and $\mathbb{T}^{-}$ by the real 7-real-dimensional cone $\mathbb{N}$. Note that if $s \neq 0$, the <u>real</u> scaling of Z^{α} is fixed by the value of s. For "classical photons" we have $s = \pm\hbar$, so it follows that they may be precisely represented by the points of $\mathbb{PT}^{\pm}$, as is required for our "second physical approximation".

In order to make detailed calculations within the twistor formalism, it is convenient to adopt the 2-spinor notation according to which a vector V^{a}, with standard orthonormal components $(V^{0},V^{1},V^{2},V^{3})$, corresponds to the spinor $V^{AA'}$, whose standard components are given by

$$\begin{pmatrix} V^{00'} & V^{01'} \\ V^{10'} & V^{11'} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} V^{0}+V^{1} & V^{2}+iV^{3} \\ V^{2}-iV^{3} & V^{0}-V^{1} \end{pmatrix}. \qquad 2.6$$

The <u>reality</u> of V^{a} corresponds to the <u>Hermiticity</u> of this (2×2) matrix. Furthermore, V^{a} is complex null, if and only if $V^{AA'}$ has the form $V^{AA'} = \xi^{A}\eta^{A'}$, and real future-pointing null whenever $V^{AA'} = \xi^{A}\bar{\xi}^{A'}$. The metric tensors g_{ab}, g^{ab} have components diag(1,-1,-1,-1) and take the spinor form $g_{AA'BB'} = \epsilon_{AB}\epsilon_{A'B'}$, $g^{AA'BB'} = \epsilon^{AB}\epsilon^{A'B'}$ where ϵ_{AB}, $\epsilon_{A'B'}$, ϵ^{AB}, $\epsilon^{A'B'}$ all have components $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$. The skew ϵ's raise and lower indices according to $\xi_{B} = \xi^{A}\epsilon_{AB}$, $\xi^{A} = \epsilon^{AB}\xi_{B}$, $\eta_{B'} = \eta^{A'}\epsilon_{A'B'}$, $\eta^{A'} = \epsilon^{A'B'}\eta_{B'}$. Round brackets around indices denote symmetrization and square brackets, anti-symmetrization. In this notation we have for (2.2), (2.3)

$$P_{AA'} = \bar{\pi}_{A}\pi_{A'} \text{ and } M^{AA'BB'} = i\,\omega^{(A}\bar{\pi}^{B)}\,\epsilon^{A'B'} - i\,\epsilon^{AB}\bar{\omega}^{(A'}\pi^{B')} \qquad 2.7$$

where (with $\pi_{A'} \neq 0$)

$$Z^{\alpha} = (\omega^{A},\pi_{A'}),\ \bar{Z}_{\alpha} = (\bar{\pi}_{A},\bar{\omega}^{A'}). \qquad 2.8$$

The expressions (2.7) for momentum and angular momentum constitue the components of the kinematic twistor

$$A^{\alpha\beta} = 2\, Z^{(\alpha} I^{\beta)\gamma} \bar{Z}_{\gamma} \ ,$$

where

$$I^{\alpha\beta} = \begin{pmatrix} \epsilon^{AB} & 0 \\ 0 & 0 \end{pmatrix}, \quad I_{\alpha\beta} = \tfrac{1}{2}\, \epsilon_{\alpha\beta\gamma\delta}\, I^{\gamma\delta} = \begin{pmatrix} 0 & 0 \\ 0 & \epsilon^{A'B'} \end{pmatrix} \qquad 2.9$$

are the skew-symmetrical infinity twistors ($\epsilon_{\alpha\beta\gamma\delta}$ being the Levi Civita twistor $\epsilon_{\alpha\beta\gamma\delta} = \epsilon_{[\alpha\beta\gamma\delta]}$, $\epsilon_{0123} = 1$). The two twistors (2.9) are actually complex conjugates of one another according to the rules (2.1), (2.8). They serve to define the line $\mathbf{I}$ in $\mathbb{P}'\mathbb{T}$, i.e. $I^{\alpha\beta} = X^{[\alpha} Y^{\beta]}$ with $\mathbf{X}$, $\mathbf{Y}$ distinct points on $\mathbf{I}$.

We can also use twistors to describe massive particles, or systems of particles, but now several twistors $Z^{\alpha}_{1}, \ldots, Z^{\alpha}_{n}$ are used. We have, for the total momentum p_a and angular momentum M^{ab},

$$P_{AA'} = \bar{\pi}_{rA}\, \pi_{rA'} \ , \quad M^{AA'BB'} = i\, \omega_{r}^{(A} \pi_{r}^{B)} \epsilon^{A'B'} - i\, \epsilon^{AB} \bar{\omega}_{r}^{(A'} \pi_{r}^{B')} \ , \qquad 2.10$$

which provide components for the general kinematic twistor

$$A^{\alpha\beta} = 2\, Z_{r}^{(\alpha} I^{\beta)\gamma} \bar{Z}_{r\gamma} \qquad 2.11$$

(summation over $r = 1, \ldots, n$ assumed). The linear freedom in Z^{α}_{r} and $\bar{Z}_{r\alpha}$, preserving p_a, M^{ab}, is given by the n-twistor internal symmetry transformations:

$$\begin{pmatrix} Z^{\alpha}_{r} \\ \cdots \\ \bar{Z}_{r\alpha} \end{pmatrix} \mapsto \left(\begin{array}{cc:c} U_{r,s}\, \delta^{\alpha}_{\beta} & & \Lambda_{r,t}\, \bar{U}_{t,s}\, I^{\alpha\beta} \\ \hdashline \bar{\Lambda}_{r,t}\, U_{t,s}\, I_{\alpha\beta} & & \bar{U}_{r,s}\, \delta_{\alpha}^{\beta} \end{array}\right) \begin{pmatrix} Z^{\beta}_{s} \\ \cdots \\ \bar{Z}_{s\beta} \end{pmatrix} .$$

Here U is an arbitrary n x n unitary matrix and Λ an arbitrary n x n complex skew matrix. The group multiplication law for the n-twistor internal group is

$$(U, \Lambda) \times (\tilde{U}, \tilde{\Lambda}) = (U\tilde{U}, \Lambda + U\tilde{\Lambda}U^{T}) \ .$$

The infinitesimal generators of each group yield an algebra of "quantum numbers", where we now generalize the commutators (2.5) to

$$[Z_r^{\alpha}, Z_t^{\beta}] = 0, \quad [\bar{Z}_{r\alpha}, \bar{Z}_{t\beta}] = 0, \quad [Z_r^{\alpha}, \bar{Z}_{t\beta}] = \hbar\delta_{\beta}^{\alpha}\,\delta_{rt} \qquad 2.12$$

When substituted into (2.10), the commutators (2.12) yield the Poincaré Lie algebra in the general (massive) case. The rest-mass of the particle (or system) turns out to be

$$\left(I_{\alpha\beta}{}_{i\ j}Z^{\alpha}Z^{\beta}I^{\rho\sigma}\bar{Z}_{i\rho}\bar{Z}_{j\sigma}\right)^{1/2}$$

and the expression for the spin is somewhat similar, though more complicated. For more details see Hughston [3,13].

If we refer the kinematics to a new origin, whose position vector with respect to O is x^a, then we find

$$p_a(x) = p_a, \qquad M^{ab}(x) = M^{ab} - 2\,x^{[a}p^{b]}\ , \qquad 2.13$$

where p_a and M^{ab} denote the values at the original origin O, so these equations define fields (of a very special kind) in space-time. Correspondingly, the twistor Z^{α} determines spinor fields:

$$\pi_{A'}(x) = \pi_{A'}, \qquad \omega^{A}(x) = \omega^{A} - i\,x^{AB'}\pi_{B'}\ , \qquad 2.14$$

these yielding the correct $p_a(x)$ and $M^{ab}(x)$ above. It is easy to show that the ω-fields of this form are precisely the solutions of the twistor equation:

$$\nabla_{A'}^{(A}\,\omega^{B)}(x) = 0\ . \qquad 2.15$$

A (classical) null twistor Z^α satisfies $\bar{Z}^\alpha \bar{Z}_\alpha = 0$ and (provided that $\pi_{A'} \neq 0$) defines a null straight line in space-time as the locus of x^a for which $\omega^A(x) = 0$, i.e. for which

$$\omega^A = ix^{AB'}\pi_{B'}\ , \qquad 2.16$$

and we say that the point x and the twistor Z^α are incident whenever this equation holds. Any real null straight line in $\mathbb{M}^I$ is obtainable in this way (and the generators of $\mathscr{I}$ arise when $\pi_{A'} = 0$). This gives us the "first physical approximation" of §1. (See fig. 2.)

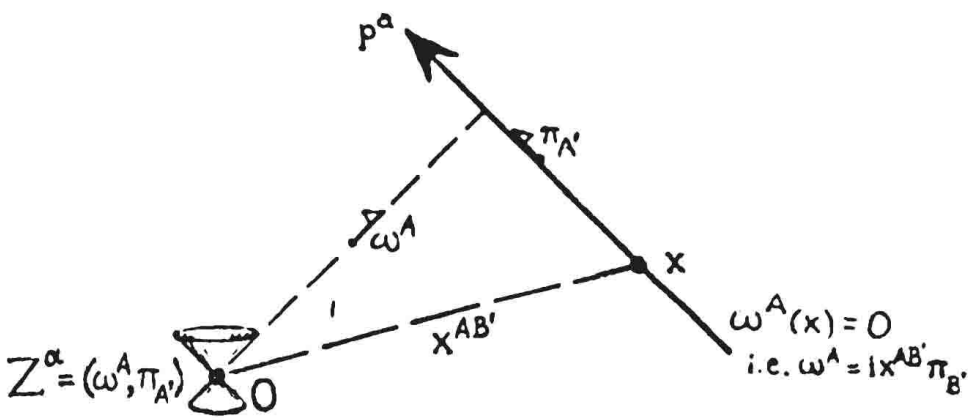


Figure 2:
The twistor description of a null line in $\mathbb{M}^I$.

In complexified Minkowski space $\mathbb{CM}$, $x^{AB'}$ need not be Hermitian and the complex solutions of (2.16) in x, for fixed ω and π - where Z^α need not now be null - define an α-plane Z, which is a 2-complex-dimensional totally null

plane in $\mathbb{CM}$. All the tangent vectors to an α-plane have the (spinor) form $\eta^{A}\pi^{A'}$, where π is fixed an η varies. These vectors are all null and mutually orthogonal. There is an important geometrical correspondence between $\mathbb{CM}$ and $\mathbb{PT}$, namely: α-planes in $\mathbb{CM}$ correspond to points in $\mathbb{PT}$; points in $\mathbb{CM}$ correspond to projective lines in $\mathbb{PT}$; two points in $\mathbb{CM}$ are null separated if and only if the corresponding lines in $\mathbb{PT}$ meet. The incidence relation $\omega^{A} = i\, x^{AB'}\pi_{B}$; between a point of x of $\mathbb{CM}$ and a twistor $Z^{\alpha} = (\omega^{A}, \pi_{A'})$ is interpreted in $\mathbb{CM}$ as the point x lying on the α-plane Z, while in $\mathbb{PT}$ it is interpreted as a line $\mathbf{X}$ passing through a point $\mathbf{Z}$. (The line $\mathbf{X}$ in $\mathbb{PT}$ corresponding to a point x of $\mathbb{CM}$ is frequently denoted L_{x}.) The space $\mathbb{CM}$ considered here is the compactified complex Minkowski space. It contains a complex light cone at infinity denoted $\mathbb{C}\mathscr{I}$ whose α-planes correspond to the points of $\mathbf{I}$, the line $\mathbf{I}$ itself being represented in $\mathbb{CM}$ by the vertex of $\mathbb{C}\mathscr{I}$. The points of $\mathbb{C}\mathscr{I}$ correspond to lines in $\mathbb{PT}$ meeting $\mathbf{I}$.

Complex conjugation in $\mathbb{CM}$ induces a duality correspondence on $\mathbb{PT}$ whereby points and planes are interchanged, this being the (projective) correspondence $Z^{\alpha} \longleftrightarrow \bar{Z}_{\alpha}$ which interchanges twistors (representing points) with dual twistors (representing planes). This correspondence sends a line in $\mathbb{PT}$ (thought of as a point locus) to another line in $\mathbb{PT}$ (thought of as the axis of a pencil of planes) - the complex conjugate line. Only the lines in $\mathbb{PN}$ are their own complex conjugates, these corresponding to the real points of $\mathbb{CM}$. Lines in $\mathbb{PT}$, not in $\mathbb{PN}$, are of five different kinds according as they lie entirely in $\mathbb{PT}^{+}$, or in the closure $\overline{\mathbb{PT}}^{+}$ $(= \mathbb{PT}^{+} \cup \mathbb{PN})$ but not in $\mathbb{PT}^{+}$, or meet all three of $\mathbb{PT}^{+}$ and $\mathbb{PN}$ and $\mathbb{PT}^{-}$,

or lie in the closure $\overline{\mathbb{P}\mathbb{T}^-}$ but not in $\mathbb{P}\mathbb{T}^-$, or lie entirely in $\mathbb{P}\mathbb{T}^-$. The corresponding points of $\mathbb{CM}$ have position vectors whose imaginary parts as past-timelike, past-null, spacelike, future-null, or future-timelike, respectively. Thus, in particular, the <u>forward tube</u> (points of $\mathbb{CM}$ with past-timelike imaginary parts) corresponds to (lines in) $\mathbb{P}\mathbb{T}^+$.

The unimodular linear transformations of $\mathbb{T}$ which preserve the twistor norm $z^\alpha \bar{z}_\alpha$, and so preserve the division of $\mathbb{T}$ into $\mathbb{T}^+$, $\mathbb{N}$ and $\mathbb{T}^-$, constitute a realization of the group SU(2,2). This induces a group of projective transformations of $\mathbb{PT}$ for which $\mathbb{PN}$ is invariant and, consequently, a group of point transformations of (compactified) Minkowski space to itself - in fact, the <u>restricted conformal group</u>, to which SU(2,2) is 4 - 1 related. To express Poincaré invariance, we restrict to the subgroup of SU(2,2) for which $I^{\alpha\beta}$ (and therefore also $I_{\alpha\beta}$) is invariant.

Note that the following objects on $\mathbb{T}$ are holomorphic and (with respect to $\mathbb{M}$) conformally invariant:

$$T = z^\alpha \frac{\partial}{\partial z^\alpha} = \text{Euler vector field (homogeneity operator)};$$

$$d^4z = \frac{1}{24}\epsilon_{\alpha\beta\gamma\delta} dz^\alpha \wedge dz^\beta \wedge dz^\gamma \wedge dz^\delta = dz^0 \wedge dz^1 \wedge dz^2 \wedge dz^3 = \frac{1}{4} d(\mathfrak{D}z),$$

$$\mathfrak{D}z = \frac{1}{6} \epsilon_{\alpha\beta\gamma\delta} z^\alpha \, dz^\beta \wedge dz^\gamma \wedge dz^\delta = T \lrcorner (d^4z);$$

while the following are holomorphic and merely Poincaré invariant:

$$d^2z = d^2\pi = \frac{1}{2} I_{\alpha\beta} \, dz^\alpha \wedge dz^\beta = \frac{1}{2} \epsilon^{A'B'} \, d\pi_{A'} \wedge d\pi_{B'}$$

$$= dz^2 \wedge dz^3 = d\pi_{0'} \wedge d\pi_{1'} = \frac{1}{2} d(\Delta z),$$

$$\Delta Z = I_{\alpha\beta}\, Z^{\alpha} dZ^{\beta} = \epsilon^{A'B'}\, \pi_{A'} d\pi_{B'}$$

$$= \pi_{0'} d\pi_{1'} - \pi_{1'} d\pi_{0'} = Z^2 dZ^3 - Z^3 dZ^2 = T\rfloor(d^2 Z),$$

$$\epsilon^{AB}\, \frac{\partial}{\partial \omega^A} \otimes \frac{\partial}{\partial \omega^B},$$

and finally the quantities $Z^{\alpha}\bar{Z}_{\alpha}$, $iZ^{\alpha} d\bar{Z}_{\alpha}$ and $i dZ^{\alpha} \wedge d\bar{Z}_{\alpha}$ are conformally invariant but not holomorphic.

§3. Massless Fields

We saw from (2.5) that Z^{α} and $\bar{Z}_{\alpha}$ are to be regarded as canonically conjugate variables. Thus, a twistor wave-function (for a one-twistor system), in the $\mathbb{T}$-representation, is a function $f(Z^{\alpha})$ "independent of $\bar{Z}_{\alpha}$", i.e. holomorphic in Z^{α}. Then the "operator Z^{α}" is represented simply as "multiplication by the twistor variable Z^{α}", while the "operator $\bar{Z}_{\alpha}$" is represented by "$-\hbar\partial/\partial Z^{\alpha}$". Now, by (2.4), the helicity operator becomes:

$$\hbar^{-1} s = -\tfrac{1}{2} Z^{\alpha}\, \partial/\partial Z^{\alpha} - 1 = -\tfrac{1}{2} T - 1. \qquad 3.1$$

In the $\mathbb{T}^*$-representation (i.e. in terms of dual twistor variables) we can use $\check{f}(\bar{Z}_{\alpha})$ "independent of Z^{α}", i.e. anti-holomorphic in Z^{α}. The, in this case, the "operator Z^{α}" is represented by "$\hbar\partial/\partial\bar{Z}_{\alpha}$" and the "operator $\bar{Z}_{\alpha}$" by "multiplication by the variable $\bar{Z}_{\alpha}$". The helicity operator becomes:

$$\hbar^{-1} s = \tfrac{1}{2} \bar{Z}_{\alpha}\, \partial/\partial\bar{Z}_{\alpha} + 1 = \tfrac{1}{2} \bar{T} + 1. \qquad 3.2$$

Thus, whichever representation is used, pure helicity states are described by homogeneous twistor functions.

For definiteness, let us stick to the $\mathbb{T}$-representation. Then the right-handed and left-handed parts of the photon

are described by twistor functions of respective homogeneities -4 and 0, while the right- and left-handed parts of the (linearized) graviton are described by functions of respective homogeneities -6 and +2. For the scalar wave equation the homogeneity is -2, while for neutrinos and antineutrinos, -1 and -3. For a free wave-function in x-space, we have holomorphicity in the forward tube ($\mathrm{Im}(x^a)$ past-timelike), which corresponds, in $\mathcal{Z}$-space, to holomorphicity in $\mathbb{T}^+$ - or in $\mathbb{P}\mathbb{T}^+$. But (apart from polynomials, which are trivial in this context) there are no homogeneous functions that are globally holomorphic on $\mathbb{T}^+$. What we need is something a little more subtle. We can cover $\mathbb{T}^+$ (or $\mathbb{P}\mathbb{T}^+$) with open sets, say $\mathcal{U}_1$, $\mathcal{U}_2$, so that there exist such functions on the intersection $\mathcal{U}_1 \cap \mathcal{U}_2$. To obtain a space-time wave-function, we perform a contour integral within the complex line ($\mathbb{C}\mathbb{P}^1$) in $\mathbb{P}\mathbb{T}^+$ that represents the point in the forward tube at which the wave function is to be evaluated (see fig. 3).

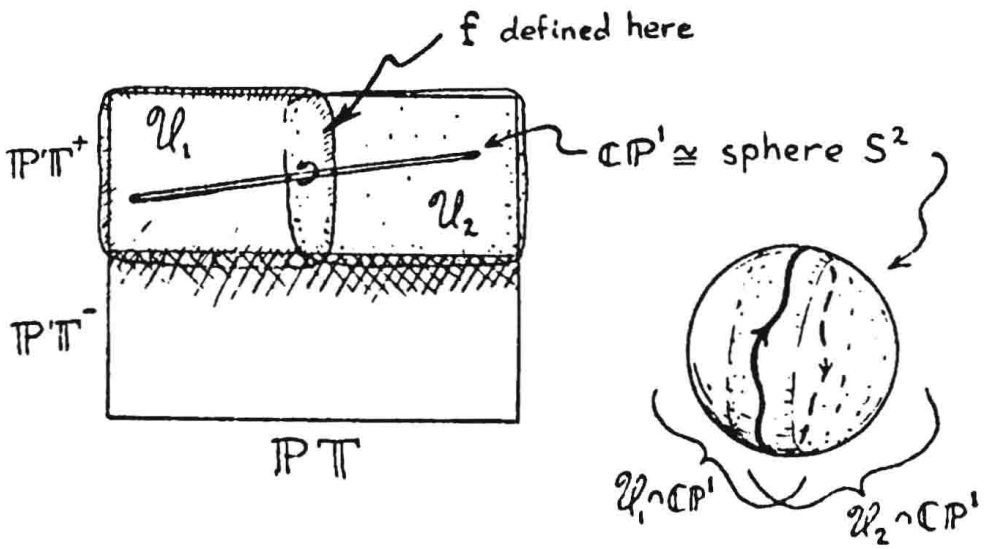


Figure 3:
A space-time wave function is
obtained by contour integration in twistor space

Then it turns out that <u>the massless free-field equations are automatically satisfied</u>:

<u>case s = 0 (hom. = -2)</u>

$$\psi(x) = \frac{1}{(2\pi i)^2}\oint f(Z)\,d^2Z \quad (Z \text{ incident with } x)$$

$$= \frac{1}{(2\pi i)^2}\oint f(ix^{AA'}\pi_{A'}, \pi_{A'})\,d^2\pi = \frac{1}{2\pi i}\oint f(\ldots)\,\Delta Z$$

$$\Rightarrow \quad \Box\psi = 0\ ;$$

<u>case s = ½nħ > 0 (hom. = -n-2)</u>

$$\psi_{A'B'\ldots L'}(x) = \frac{1}{(2\pi i)^2}\oint \pi_{A'}\pi_{B'}\ldots\pi_{L'}\, f(ix^{AA'}\pi_{A'}, \pi_{A'})\,d^2\pi$$

$$\Rightarrow \psi_{A'B'\ldots L'} = \psi_{(A'B'\ldots L')},\ \nabla^{AA'}\psi_{A'B'\ldots L'} = 0\ ;$$

<u>case s = -½nħ < 0 (hom. = n-2)</u>

$$\psi_{AB\ldots L}(x) = \frac{1}{(2\pi i)^2}\oint \frac{\partial}{\partial\omega^A}\frac{\partial}{\partial\omega^B}\ldots\frac{\partial}{\partial\omega^L} f(ix^{AA'}\pi_{A'}, \pi_{A'})\,d^2\pi$$

$$\Rightarrow \psi_{AB\ldots L} = \psi_{(AB\ldots L)},\ \nabla^{AA'}\psi_{AB\ldots L} = 0\ .$$

In these equations $\nabla_{AA'}$ stands for $\partial/\partial x^{AA'}$, i.e. it is the spinor translation of $\nabla_a = \partial/\partial x^a$, while $\Box = \nabla_a\nabla^a$.

The special cases s = ±½ħ give the Dirac-Weyl neutrino equation. When s = ±ħ we get a (complex) Maxwell field $F_{ab}(= -F_{ab})$:

$$F_{AA'BB'} = \underbrace{\psi_{AB}\,\epsilon_{A'B'}}_{\substack{\text{anti-self-dual}\\ \text{(left-handed)}}} + \underbrace{\epsilon_{AB}\,\tilde{\psi}_{A'B'}}_{\substack{\text{self-dual}\\ \text{(right-handed)}}}$$

the Maxwell equations $\nabla^a F_{ab} = 0$, $\nabla_{[a}F_{bc]} = 0$ becoming $\nabla^{AA'}\psi_{AB} = 0$, $\nabla^{AA'}\tilde{\psi}_{A'B'} = 0$. When s = ±2ħ we get a (complex)

Einstein curvature field K_{abcd}:

$$K_{AA'BB'CC'DD'} = \underbrace{\Psi_{ABCD}\epsilon_{A'B'}\epsilon_{C'D'}}_{\text{anti-self-dual (left-handed)}} + \underbrace{\epsilon_{AB}\epsilon_{CD}\tilde{\Psi}_{A'B'C'D'}}_{\text{self-dual (right-handed)}}$$

We have $K_{abcd} = K_{[cd][ab]}, K_{[abc]d} = 0$, $K^{a}{}_{bad} = 0$. The "Bianchi identity" $\nabla^{a}K_{abcd} = 0$, or equivalently $\nabla_{[a}K_{bc]de} = 0$, becomes $\nabla^{AA'}\Psi_{ABCD} = 0$ and $\nabla^{AA'}\Psi_{A'B'C'D'} = 0$. In fact these constructions yield the general positive frequency fields for each helicity.

We are now in a position to see how a twistor P^{α} and a dual twistor Q_{α} can play respective roles as helicity raising and helicity lowering operators, in accordance with the "fourth physical approximation" mentioned in §1. For given a twistor function $f(Z)$, of homogeneity degree $-r-2$ we can construct two new ones

$$g = P^{\alpha}\,\partial f/\partial Z^{\alpha} \text{ and } h = Q_{\alpha}Z^{\alpha}f\ ,$$

the homogeneity degree having been respectively reduced and increased by unity in the two cases; i.e. the helicity respectively increased and reduced by $\frac{1}{2}\hbar$. When g is substituted into the appropriate contour integral expression, we obtain

$$-iP^{M}(x)\nabla_{MM'}\Psi_{A'B'\ldots L'} - (r+1)P_{(M'}\Psi_{A'B'\ldots L')} \qquad (r \geq 0)$$

$$\text{or} \qquad -P^{L}(x)\Psi_{AB\ldots KL}\ ; \qquad (r < 0)$$

when h is substituted, we obtain

$$Q^{L'}(x)\Psi_{A'B'\ldots K'L'} \qquad (r > 0)$$

or

$$-iQ^{M'}(x)\nabla_{MM'}\Psi_{AB\ldots L} - (r-1)Q_{(M}\Psi_{AB\ldots L)} \qquad (r \leq 0)\ .$$

Here, we have

$$P^{A}(x) = P^{A} - ix^{AB'}P_{B'} \quad \text{with} \quad P^{\alpha} = (P^{A}, P_{A'})$$

and

$$Q^{A'}(x) = Q^{A'} + ix^{BA'}Q_{B} \quad \text{with} \quad Q_{\alpha} = (Q_{A}, Q^{A'})$$

The satisfaction of the field equations, in each case, arises from the twistor equation (2.15).

There is much freedom in the choice of twistor function f, yielding a given $\psi_{\ldots}$, according to the prescription of figure 3. For if f is replaced

$$f \to f + q_1 + q_2$$

where q_1 is holomorphic on $\mathcal{U}_1$ and q_2 holomorphic on $\mathcal{U}_2$, then the result of the contour integration will be the same as before, the contributions from q_1 and q_2 vanishing because the contour "slips off" on one side or the other. Furthermore, the domain $\mathcal{U}_1 \cap \mathcal{U}_2$ on which f is defined is to a large degree arbitrary. The complete freedom involved is somewhat complicated and is best understood in terms of the ideas of sheaf cohomology[14]. The discussion of twistor functions from this point of view is beyond the scope of this article, but detailed accounts can be found elsewhere[15,16].

One significant feature of this description may be remarked upon here. We have a description of <u>linear</u> massless fields according to which twistor functions are defined on intersections of overlapping open sets. There are also various <u>non-linear</u> versions of the same type of prodecure, in which the twistor functions are replaced by suitable <u>transition functions</u>, i.e. by functions which effectively describe how two abstract (twistor) spaces are to be "glued

together". In many cases, twistor functions may be viewed essentially as "infinitesimal transition functions". Generalizing to the case of "finite" transition functions it has been possible to reduce the anti-self-dual Einstein[2,11] and Yang-Mills[1,12] equations effectively to twistor holomorphicity. These, and many related ideas, are the subject of much active work[16,17]. It does not seem to be unreasonable to expect that physical interactions, generally, should eventually be expressible in terms of similar constructions.

References

1. R.S. Ward, S.H.E.P.
2. K.P. Tod, S.H.E.P.
3. L.P. Hughston, S.H.E.P.
4. A. Hodges & S. Huggett, S.H.E.P.
5. R. Penrose, J. Math. Phys. 8 (1967) 345.
6. R. Penrose, Repts. Math. Phys. 12 (1977) 65.
7. R. Penrose & M.A.H. MacCallum, Phys. Rep. 6C, (1973) 241.
8. R. Penrose in Quantum Gravity, an Oxford Symposium, eds. C.J. Isham, R. Penrose and D.W. Sciama (Oxford University Press 1975).
9. R. Penrose & R.S. Ward, General Relativity and Gravitation, One Hundred Years after Einstein, Vol. 2, ed. A. Held (Plenum 1980).
10. R. Penrose, J. Math. Phys. 10 (1969) 38.
11. R. Penrose, G.R.G. 7 (1976) 31.
12. R.S. Ward, Phys. Lett. 59B (1977) 81.
13. L.P. Hughston, Twistors and Particles (Springer 1979).
14. J. Morrow and K. Kodaira, Complex Manifolds (Holt, Rinehart and Winston, 1971).
15. M.G. Eastwood, R. Penrose and R.O. Wells, Jr., Comm. Math. Phys. to appear (1980).

16. L.P. Hughston and R.S. Ward (eds.), Advances in Twistor Theory (Pitman APP, 1979).

17. D.E. Lerner and P.D. Sommers (eds.), Complex Manifold Techniques in Theoretical Physics (Pitman APP 1979).

120
A BRIEF OUTLINE OF TWISTOR THEORY

Cosmology and gravitation; spin, torsion, rotation and supergravity. Eds P. G. Bergmann and V. de Sabbata, pp. 287–316, Plenum, *New York,* 1980

664 A brief outline of twistor theory. *Cosmology and gravitation; spin, torsion, rotation and supergravity.* Eds P. G. Bergmann and V. de Sabbata, pp. 287–316. Plenum, *New York,* 1980

A BRIEF OUTLINE OF TWISTOR THEORY

Roger Penrose

Mathematical Institute
Oxford, OX1 3LB
England

1. Motivation

The normal space-time formalism for physics has proved remarkably successful. There are, nevertheless, several reasons for believing that, at a certain level of understanding, these space-time concepts will give way to others that are even more basic and fundamental. In the first place, the concept of space-time "point" is not very directly physical. For example, the particles that constitute matter as we know it are, even at the classical level, one-dimensional objects (world-lines) rather than zero-dimensional. Then, when quantum theory enters the picture, the uncertainty princple serves to obscure even further the relation between physical particles and space-time points. Do points really "exist" in a precisely defined sense at 10^{-13}cm? The question becomes even more pertinent at the quantum gravity level of 10^{-33}cm.

It is worth while to point out, also, the conflict of formalisms between that of relativity, which uses a four-dimensional (+---) pseudo-Riemannian real space, whose basic constitutes are "events", and that of quantum mechanics, which uses a

positive-definite complex Hilbert space (of unspecified dimension) whose basic constituents are "states." What would seem desirable is a formalism which treats the space-time and quantum aspects of physics in a more unified way. Twistor theory, to a large extent achieves this. It uses the complex field $\mathbb{C}$ as its basic number system and the normal real space-time geometry may be derived. Furthermore, the space-time geometry that emerges is four-dimensional, with the correct (+---) signature, so one may say that the twistor formalism is specially tailored to the particular space-time structure needed for physics rather than being a general mathematical formalism.

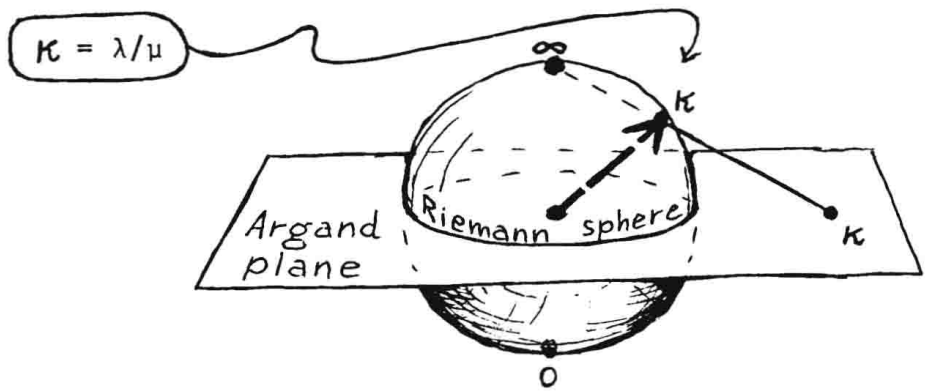
In the first instance, twistor theory merely provides a reformulation of the basic concepts and, in this sense is not a "new" theory. But it suggests different ways to proceed and new hypotheses suggest themselves at a later stage. (A possible parallel would be with Cartan's reformulation of Newtonian gravitation as a space-time theory. In this form, the modifications needed to change the theory into full-blown general relativity would have seemed very simple and natural indeed.) One of the hopes, of course, is that by providing a new framework for the description of space-time and quantum physics some of the difficulties with existing theory, such as the infinities of quantum field theory and the interpretational problems of ordinary quantum theory may eventually be removed.

The "successes" of twistor theory, so far, have been mainly of a mathematical nature. The theory provides a nice representation of Minkowski-space geometry, of classical spinning massless particles and of free massless fields. It provides a programme for studying massive particles and their internal symmetry groups. Moreover, it provides a general method for solving the non-linear Einstein and Yang-Mills field equations in the (anti-) self-dual case. There is some prospect that this

(anti-) self-dual restriction can be ultimately removed.

The complex numbers of quantum linear supersition have a relation to the geometry of space-time, and it is this that may be regarded as the starting point of the twistor approach. Consider, first, a non-relativistic particle of spin $\frac{1}{2}$. The complex linear combinations of the two states "spin up" and "spin down" provide the states of spin in all other spatial directions. This relates the sphere of directions in physical three-dimensional space to the Riemann sphere of ratios of pairs of complex numbers:

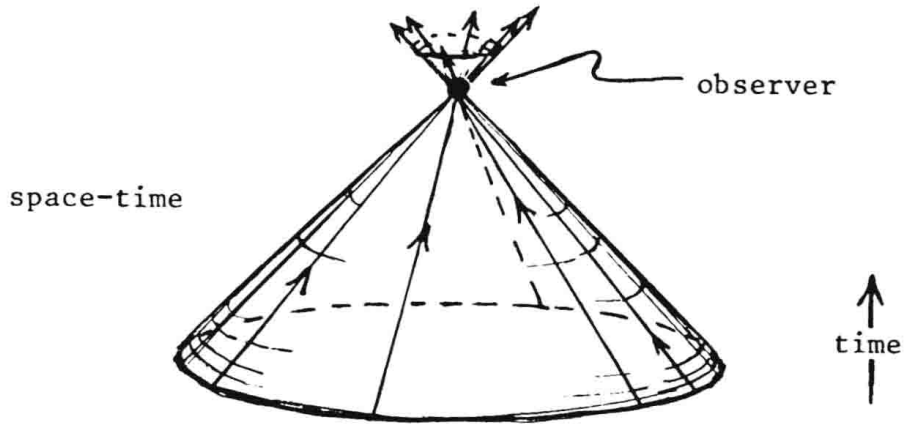
λ ↑ + μ ↓ = ↗ (spin $\frac{1}{2}$).



The rotations of space correspond, according to the 2-1 map $SU(2) \to SO(3)$, to the unitary fractional linear transformations. Dropping the unitary condition we get the relativistic version of this, which is the 2-1 map $SL(2,\mathbb{C}) \to O_+^{\uparrow}(1,3)$. The holomorphic transformations

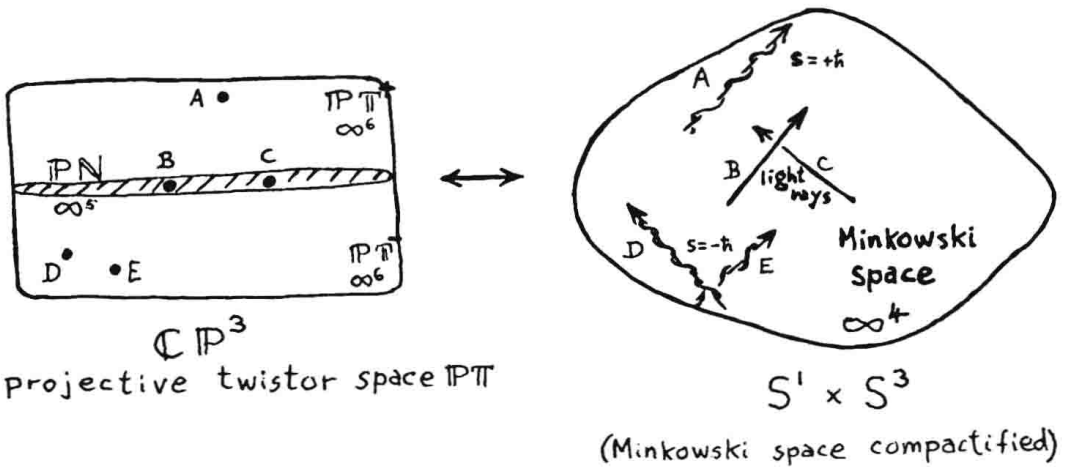
$$\kappa \longmapsto \frac{\alpha\kappa+\beta}{\gamma\kappa+\delta}$$

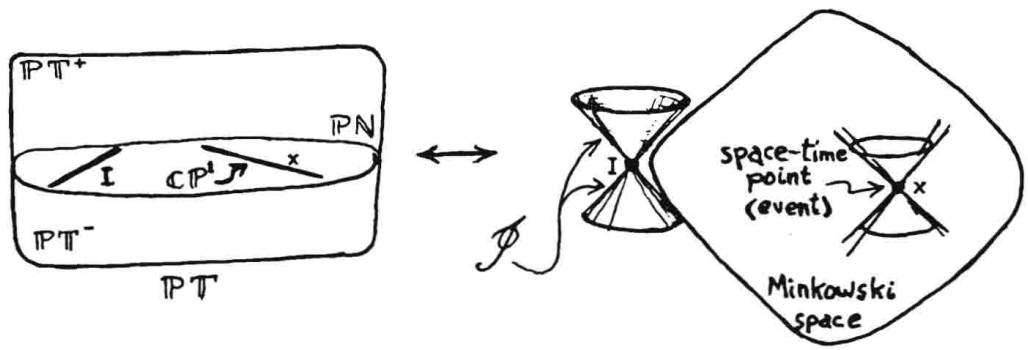
of the Riemannian sphere of κ to itself thus provide the elements of the restricted Lorentz group. In terms of space-time geometry, this Riemann sphere is the celestial sphere of an observer, namely the family of all light-rays through a point of Minkowski space-time:



Thus, as an abstract manifold, the space of light-rays through a fixed point is, in a natural way, a complex 1-dimensional manifold (namely the Riemann sphere). Consider, now, the space of all light-rays (null straight lines) in Minkowski space, not merely the ones through a fixed point. This is a real 5-dimensional space and so cannot, as a whole, be a complex manifold. Yet, remarkably, it is not far from being one. If we consider not just light-rays but photons with spin and energy, then fixing the helicity s to take the value $s = +\hbar$ we obtain a real 6-dimensional manifold whose structure is, indeed, that of a complex 3-dimensional manifold. As we shall see shortly, this manifold, which we label $\mathbb{P}\mathbb{T}^+$, is one (open) half of a complex projective 3-space $\mathbb{C}\mathbb{P}^3$ (labelled $\mathbb{P}\mathbb{T}$) the other (open) half $\mathbb{P}\mathbb{T}^-$ representing the $s = -\hbar$ photons. The real 5-dimensional boundary $\mathbb{P}\mathbb{N}$ between the two can be thought of as the space of light-rays in Minkowski space, $\mathbb{P}\mathbb{N}$ containing a special projective line ($\mathbb{C}\mathbb{P}^1$) labelled I, that represents the family of idealized light-rays that generate the light-cone $\mathscr{I}$ at infinity.

 A brief outline of twistor theory. *Cosmology and gravitation; spin, torsion, rotation and supergravity.* Eds P. G. Bergmann and V. de Sabbata, pp. 287–316. Plenum, *New York*, 1980





2. Explicit Description of a Twistor

A twistor Z^α is an element of a complex 4-dimensional vector space $\mathbb{T}$ (twistor space). The components $(z^0,z^1,z^2,z^3) \in \mathbb{C}^4$ of Z^α, together with the components

$$(\bar{z}_0,\bar{z}_1,\bar{z}_2,\bar{z}_3) = (\overline{z^2},\overline{z^3},\overline{z^0},\overline{z^1})$$

of its complex conjugate dual twistor $\bar{Z}_\alpha$, determine the components of the momentum p^a and angular momentum M^{ab} of a massless free particle (in Minkowski space) according to:

$$P^0 = \frac{1}{\sqrt{2}}(z^3\bar{z}_1 + z^2\bar{z}_0),\ P^1 = \frac{1}{\sqrt{2}}(z^3\bar{z}_1 - z^2\bar{z}_0)$$

$$P^2 = \frac{1}{\sqrt{2}}(-z^2\bar{z}_1 - z^3\bar{z}0),P^3 = \frac{i}{\sqrt{2}}(z^2\bar{z}_1 - z^3\bar{z}_0)$$

$$M^{01} = -M^{10} = \tfrac{i}{2}(Z^0\bar{Z}_0 - Z^1\bar{Z}_1 - Z^2\bar{Z}_2 + Z^3\bar{Z}_3)$$
$$M^{02} = -M^{20} = \tfrac{i}{2}(Z^0\bar{Z}_1 + Z^1\bar{Z}_0 - Z^2\bar{Z}_3 - Z^3\bar{Z}_2)$$
$$M^{03} = -M^{30} = \tfrac{1}{2}(Z^0\bar{Z}_1 - Z^1\bar{Z}_0 - Z^2\bar{Z}_3 + Z^3\bar{Z}_2)$$
$$M^{12} = -M^{21} = \tfrac{i}{2}(Z^0\bar{Z}_1 - Z^1\bar{Z}_0 + Z^2\bar{Z}_3 - Z^3\bar{Z}_2)$$
$$M^{32} = -M^{23} = \tfrac{1}{2}(Z^0\bar{Z}_0 - Z^1\bar{Z}_1 + Z^2\bar{Z}_2 - Z^3\bar{Z}_3)$$
$$M^{13} = -M^{31} = \tfrac{1}{2}(Z^0\bar{Z}_1 + Z^1\bar{Z}_0 + Z^2\bar{Z}_3 + Z^3\bar{Z}_2)$$
$$M^{00} = M^{11} = M^{22} = M^{33} = 0.$$

In addition to the required properties $p_a p^a = 0$, $p^0 > 0$ automatically holding, the Pauli-Łubanski spin vector

$$S_a = \tfrac{1}{2}\, e_{abcd}\, p^b\, M^{cd}$$

also automatically satisfies the necessary relation

$$S_a = s\, p_a\ ,$$

where the helicity s is given by

$$s = \tfrac{1}{2}\, Z^\alpha \bar{Z}_\alpha\ .$$

The twistor Z^α is determined up to the phase transformation $Z^\alpha \longmapsto e^{i\theta} Z^\alpha$ (θ real) by the p^a, M^{ab} of a massless particle and, conversely, every twistor (except those for which $Z^2 = Z^3 = 0$) is related to a (finite) massless particle in this way.

Projective twistor space $\mathbb{PT}$ is the space of proportionality classes of non-zero twistors. The standard coordinates for $\mathbb{PT}$ are the three independent complex ratios $Z^0{:}Z^1{:}Z^2{:}Z^3$. $\mathbb{PT}$ is divided into two complex 3-manifolds $\mathbb{PT}^+$ (given by $Z^\alpha\bar{Z}_\alpha > 0$) and $\mathbb{PT}^-$ (given by $Z^\alpha\bar{Z}_\alpha < 0$), by their common 5-real-dimensional boundary $\mathbb{PN}$ (given by $Z^\alpha\bar{Z}_\alpha = 0$). Similarly, twistor space $\mathbb{T}$ is divided into $\mathbb{T}^+$ and $\mathbb{T}^-$ by the 7-real-dimensional cone $\mathbb{N}$.

In order to make detailed calculations within the twistor

 A brief outline of twistor theory. *Cosmology and gravitation; spin, torsion, rotation and supergravity.* Eds P. G. Bergmann and V. de Sabbata, pp. 287–316. Plenum, *New York,* 1980

formalism, it is convenient to adopt the 2-spinor notation according to which a vector V^a, with standard orthonormal components (V^0,V^1,V^2,V^3), corresponds to the spinor $V^{AA'}$, whose standard components are given by

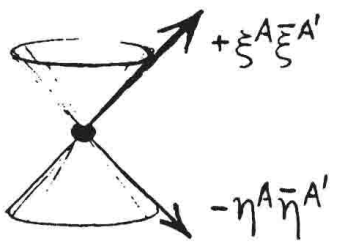
$$\begin{pmatrix} V^{00'} & V^{01'} \\ V^{10'} & V^{11'} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} V^0+V^1 & V^2+V^3 i \\ V^2-V^3 i & V^0-V^1 \end{pmatrix}.$$

The reality of V^a corresponds to the Hermiticity of this (2 x 2) matrix. Furthermore,

V^a null $\Leftrightarrow \det(V^{AA'}) = 0$

$\Leftrightarrow V^{AA'} = \xi^A \eta^{A'}$

$= \pm\xi^A \bar{\xi}^{A'}$ if V^a real.



The metric tensors g_{ab}, g^{ab} have components diag(1,-1,-1,-1) and take the spinor form $g_{AA'BB'} = \epsilon_{AB}\epsilon_{A'B'}$, $g^{AA'BB'} = \epsilon^{AB}\epsilon^{A'B'}$ where $\epsilon_{AB}, \epsilon_{A'B'}, \epsilon^{AB}, \epsilon^{A'B'}$ all have components $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$. The skew ϵ's raise and lower indices according to $\xi_B = \xi^A\epsilon_{AB}$, $\xi^A = \epsilon^{AB}\xi_B$, $\eta_{B'} = \eta^{A'}\epsilon_{A'B'}$, $\eta^{A'} = \epsilon^{A'B'}\eta_{B'}$. In this notation we have, for the momentum and angular momentum;

$$P_{AA'} = \bar{\pi}_A \pi_{A'}$$

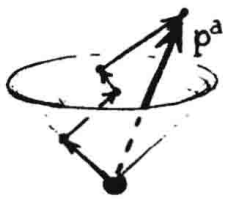
$$M^{AA'BB'} = i\,\omega^{(A}\pi^{B)}\epsilon^{A'B'} - i\,\epsilon^{AB}\bar{\omega}^{(A'}\bar{\pi}^{B')},$$

where (with $\pi_{A'} \neq 0$)

$$Z^\alpha = (\omega^A, \pi_{A'})$$

$$\bar{Z}_\alpha = (\bar{\pi}_A, \bar{\omega}^{A'}).$$

We can also use twistors to describe massive particles (or systems of particles), but now more than one twistor $Z_1^\alpha,\ldots,Z_n^\alpha$ must be used. We have, for the total momentum p_a and angular momentum M^{ab},



$$P_{AA'} = \bar{\pi}_{rA}\,\pi_{rA'}$$

$$M^{AA'BB'} = i\,\omega_r^{(A}\pi_r^{B)}\epsilon^{A'B'} - i\,\epsilon^{AB}\bar{\omega}_r^{(A'}\bar{\pi}_r^{B')}$$

(summation over r = 1,...,n assumed). The linear freedom in Z^α_r and $\bar{Z}_{r\alpha}$ is given by the n-twistor internal symmetry transformations:

$$\begin{pmatrix} Z^\alpha_r \\ \hline \bar{Z}_{r\alpha} \end{pmatrix} \longmapsto \left(\begin{array}{c|c} U_{r,s}\,\delta^\alpha_\beta & \Lambda_{r,t}\,U_{t,s}\begin{pmatrix}\epsilon^{AB} & 0\\ 0 & 0\end{pmatrix} \\ \hline \bar{\Lambda}_{r,t}\,U_{t,s}\begin{pmatrix}0 & 0\\ 0 & \epsilon_{A'B'}\end{pmatrix} & \bar{U}_{r,s}\,\delta_\alpha^\beta \end{array}\right) \begin{pmatrix} Z^\beta_s \\ \hline \bar{Z}_{s\beta} \end{pmatrix}$$

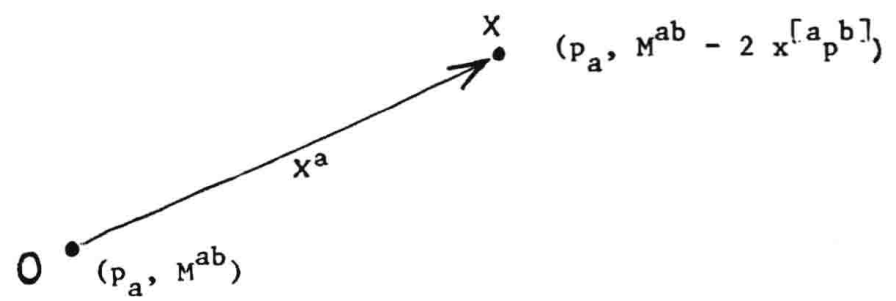
Here U is an arbitrary n x n unitary matrix and Λ an arbitrary n x n complex skew matrix. The group multiplication law for the n-twistor internal group is

$$(U, \Lambda) \times (\tilde{U}, \tilde{\Lambda}) = (U\tilde{U},\ \Lambda + U\tilde{\Lambda}U^{\mathsf{T}}) .$$

According to (naïve) twistor particle theory, the (quantized) infinitesimal generators of these groups provide an algebra of quantum numbers for elementary particles. I shall not discuss this theory here, however.

3. Change of Origin

The momentum and angular momentum of a system in special relativity have very specific transformation properties under change of origin:



A brief outline of twistor theory. *Cosmology and gravitation; spin, torsion, rotation and supergravity.* Eds P. G. Bergmann and V. de Sabbata, pp. 287–316. Plenum, *New York,* 1980

Thus, we can, if desired, think of p_a and M^{ab} as <u>fields</u> (i.e. functions of position) given by

$$p_a(x) = p_a \ ,$$
$$M^{ab}(x) = M^{ab} - 2\, x^{[a}p^{b]} \ ,$$

where p_a and M^{ab} denote the values at the original fixed origin O. Correspondingly, the twistor Z^α determines spinor fields:

$$\pi_{A'}(x) = \pi_{A'}$$
$$\omega^A(x) = \omega^A - i\, x^{AB'}\pi_{B'}$$

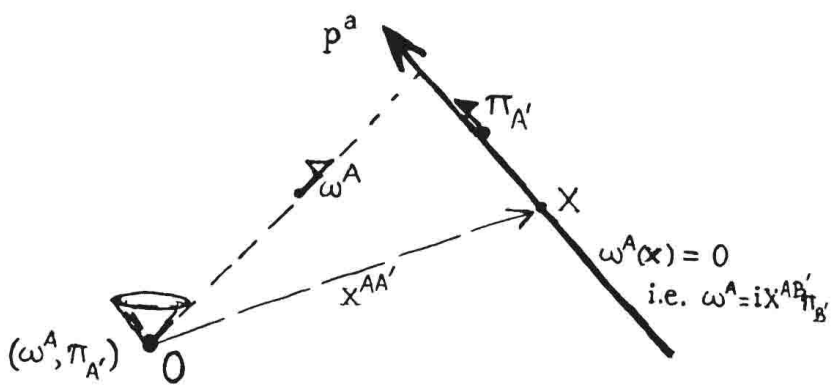
these yielding the correct $p_a(x)$ and $M^{ab}(x)$ in any one- or many-twistor system. It is easy to show that the ω-fields of this form are precisely the solutions of the twistor equation:

$$\nabla^{(A}_{A'}\, \omega^{B)}\ (x) = 0 \ .$$

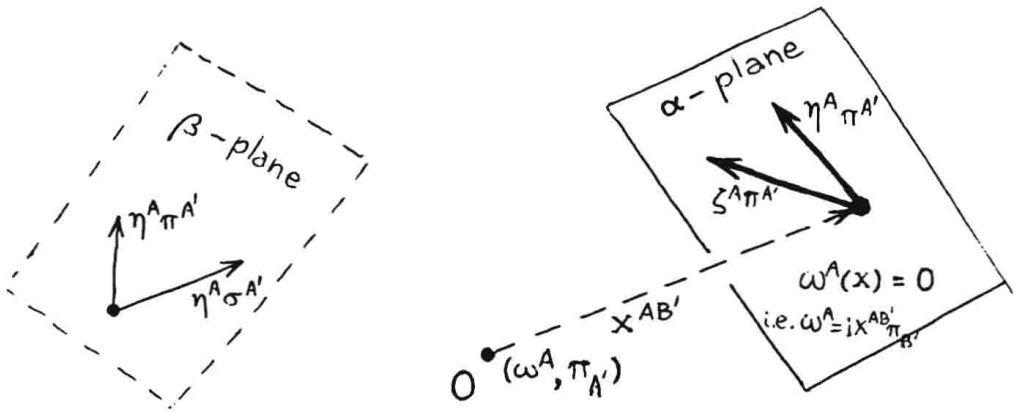
A null twistor Z^α satisfies $Z^\alpha \bar{Z}_\alpha = 0$ and (provided that $\pi_{A'} \neq 0$) defines a null straight line in space-time as the locus of x^a for which $\omega^A(x) = 0$, i.e. for which

$$\omega^A = ix^{AB'}\pi_{B'}$$

and we say that the point x and the twistor Z^α are <u>incident</u> whenever this equation holds. Any real null straight line is obtainable in this way.



In complexified Minkowski space $\mathbb{CM}$, $x^{AB'}$ need not be Hermitian and the complex solutions of the above equation in x, for fixed ω and π - where Z^{α} need not now be null - define an α-plane



which is a 2-complex-dimensional totally null plane in $\mathbb{CM}$. All the tangent vectors to an α-plane have the (spinor) form $\eta^A\pi^{A'}$, where π is fixed and η varies. These vectors are all null and mutually orthogonal. Similarly a β-plane may be defined whose tangent vectors have the form $\eta^A\pi^{A'}$ where now it is the unprimed spinor η that is fixed and π that varies. There is an important geometrical correspondence between $\mathbb{CM}$ and $\mathbb{PT}$, namely: α-planes in $\mathbb{CM}$ correspond to points in $\mathbb{PT}$; points in $\mathbb{CM}$ correspond to projective lines in $\mathbb{PT}$; β-planes in $\mathbb{CM}$ correspond to projective planes in $\mathbb{PT}$; two points in $\mathbb{CM}$ are null separated if and only if the corresponding lines in $\mathbb{PT}$ meet; a (complex) null geodesic in $\mathbb{CM}$ corresponds to a plane pencil in $\mathbb{PT}$, i.e. the system of lines through a point and lying in one plane, this point and plane in $\mathbb{PT}$ defining a unique α-plane and β-plane in $\mathbb{CM}$ through the null geodesic. The incidence relation $\omega^A = i\,x^{AB'}\pi_{B'}$, between a point x of $\mathbb{CM}$ and a twistor $Z^{\alpha} = (\omega^A, \pi_{A'})$ is interpreted in $\mathbb{CM}$ as the point x lying on the α-plane Z, while in $\mathbb{PT}$ it is interpreted as the line x passing through the point Z.

 A brief outline of twistor theory. *Cosmology and gravitation; spin, torsion, rotation and supergravity.* Eds P. G. Bergmann and V. de Sabbata, pp. 287–316. Plenum, *New York,* 1980

The space $\mathbb{CM}$ considered here is the compactified complex Minkowski space. It contains a complex light cone at infinity denoted $\mathbb{C}\mathscr{I}$. The generators of $\mathbb{C}\mathscr{I}$ correspond to the points of a special line I in $\mathbb{PT}$, the line I itself being represented in $\mathbb{CM}$ by the vertex of $\mathbb{C}\mathscr{I}$. The points of $\mathbb{C}\mathscr{I}$ correspond to lines in $\mathbb{PT}$ meeting I. The real points of $\mathbb{C}\mathscr{I}$, i.e. the points of $\mathscr{I}$, correspond to lines in $\mathbb{PN}$ meeting I. The finite real points of Minkowski space correspond to lines in $\mathbb{PN}$ not meeting I. A complex (finite) point in $\mathbb{CM}$ corresponds to a general line in $\mathbb{PT}$ not meeting I. Complex conjugation in $\mathbb{CM}$ interchanges α-planes and β-planes. Thus, this induces a duality correspondence on $\mathbb{PT}$ whereby points and planes are interchanged. In fact this is simply the (projective) correspondence $Z^\alpha \longleftrightarrow \bar{Z}_\alpha$ which interchanges twistors with dual twistors. This correspondence sends a line in $\mathbb{PT}$ (thought of as a point locus) to another line in $\mathbb{PT}$ (thought of as the axis of a pencil of planes) - the complex-conjugate line. Only the lines in $\mathbb{PN}$ are their own complex conjugates, these corresponding to the real points of $\mathbb{CM}$. Lines in $\mathbb{PT}$, not in $\mathbb{PN}$, are of five different kinds according as they lie entirely in $\mathbb{PT}^+$, or in the closure $\overline{\mathbb{PT}}^+$ (= $\mathbb{PT}^+ \cup \mathbb{PN}$) but not in $\mathbb{PT}^+$, or meet all three of $\mathbb{PT}^+$ and $\mathbb{PN}$ and $\mathbb{PT}^-$, or lie in the closure $\overline{\mathbb{PT}}^-$ but not in $\mathbb{PT}^-$, or lie entirely in $\mathbb{PT}^-$. The corresponding points of $\mathbb{CM}$ have position vectors whose imaginary parts are past-timelike, past-null, spacelike, furture-null, or future-timelike, respectively. Thus, in particular, the forward tube (points of $\mathbb{CM}$ with past-timelike imaginary parts) corresponds to (lines in) $\mathbb{PT}^+$.

The (unimodular) linear transformations of $\mathbb{T}$ which preserve the twistor norm $Z^\alpha \bar{Z}_\alpha$, and so preserve the division of $\mathbb{T}$ into $\mathbb{T}^+$, $\mathbb{N}$ and $\mathbb{T}^-$, consitute a realization of the group SU(2,2). This induces a group of projective transformations of $\mathbb{PT}$ for which $\mathbb{PN}$ is invariant and, consequently, a group of

point transformations of (compactified) Minkowski space to itself - in fact, the restricted conformal group, to which SU(2,2) is 4 - 1 related.

4. Poincaré Invariance

To express Poincaré invariance in $\mathbb{CM}$, and not just conformal invariance, we need to preserve more of the structure of $\mathbb{T}$. SU(2,2) preserves the twistor norm $Z^{\alpha}\bar{Z}_{\alpha}$ and, for unimodularity, the twistor Levi Civita symbols $\epsilon_{\alpha\beta\gamma\delta}$ and $\epsilon^{\alpha\beta\gamma\delta}$. The subgroup of SU(2,2) corresponding to the restricted Poincaré group also preserves the infinity twistors

$$I^{\alpha\beta} = \begin{pmatrix} \epsilon^{AB} & 0 \\ 0 & 0 \end{pmatrix}, \quad I_{\alpha\beta} = \begin{pmatrix} 0 & 0 \\ 0 & \epsilon^{A'B'} \end{pmatrix}.$$

These twistors serve to define the line I in $\mathbb{PT}$.

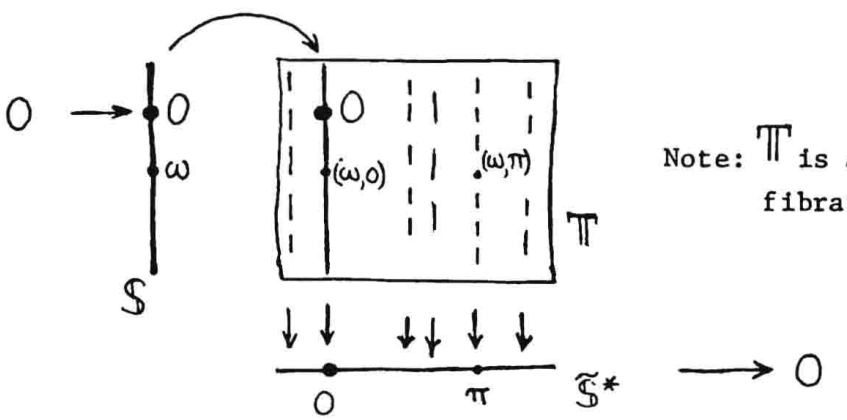
An alternative way to think of this additional Poincaré-invariant structure for $\mathbb{T}$ is as the exact sequence

$$0 \longrightarrow \mathbb{S} \longrightarrow \mathbb{T} \longrightarrow \tilde{\mathbb{S}}^* \longrightarrow 0$$

$$\omega^A \longmapsto (\omega^A, 0)$$

$$(\omega^A, \pi_{A'}) \longmapsto \pi_{A'} \ .$$

Here $\mathbb{S}$ denotes unprimed spin-space, $\tilde{\mathbb{S}}$ denotes primed spin-space, and the arterisk means dual.



Note: $\mathbb{T}$ is a holomorphic fibration over $\tilde{\mathbb{S}}^*$

A brief outline of twistor theory. *Cosmology and gravitation; spin, torsion, rotation and supergravity.* Eds P. G. Bergmann and V. de Sabbata, pp. 287–316. Plenum, *New York,* 1980

The dual of this sequence is also, when read in reverse order, its complex conjugate, and is

$$0 \longleftarrow \mathbb{S}^* \longleftarrow \mathbb{T}^* \longleftarrow \tilde{\mathbb{S}} \longleftarrow 0$$

$$(0,\tilde{\omega}^{A'}) \longleftarrow \tilde{\omega}^{A'}$$

$$\tilde{\pi}_A \longleftarrow (\tilde{\pi}_A,\tilde{\omega}^{A'}) .$$

We have the following invariant structure defined on $\mathbb{T}$:

holomorphic, conformally invariant

$\Upsilon = Z^\alpha \dfrac{\partial}{\partial Z^\alpha}$, Euler vector field (homogeneity operator);

$$d^4Z = \frac{1}{24}\,\epsilon_{\alpha\beta\gamma\delta}\, dZ^\alpha \wedge dZ^\beta \wedge dZ^\gamma \wedge dZ^\delta$$

$$= dZ^0 \wedge dZ^1 \wedge dZ^2 \wedge dZ^3$$

$$= \frac{1}{4}\, d(\mathfrak{D} Z), \text{ complex volume 4-form;}$$

$$\mathfrak{D} Z = \frac{1}{6}\,\epsilon_{\alpha\beta\gamma\delta}\, Z^\alpha\, dZ^\beta \wedge dZ^\gamma \wedge dZ^\delta$$

$$= Z^0 dZ^1 \wedge dZ^2 \wedge dZ^3 - Z^1 dZ^0 \wedge dZ^2 \wedge dZ^3 + Z^2 dZ^0 \wedge dZ^1 \wedge dZ^3 - Z^3 dZ^0 \wedge dZ^1 \wedge dZ^2$$

$$= \Upsilon \lrcorner (d^4Z), \text{ projective volume 3-form;}$$

holomorphic, Poincaré invariant

$$d^2Z = d^2\pi = \frac{1}{2}\, I_{\alpha\beta}\, dZ^\alpha \wedge dZ^\beta = \frac{1}{2}\,\epsilon^{A'B'}\, d\pi_{A'} \wedge d\pi_{B'}$$

$$= dZ^2 \wedge dZ^3 = d\pi_{0'} \wedge d\pi_{1'}$$

$$= \frac{1}{2}\, d(\Delta Z), \text{ pull-back of "volume" (area) 2-form on } \mathbb{S}^* ;$$

$$\Delta Z = I_{\alpha\beta}\, Z^\alpha dZ^\beta = \epsilon^{A'B'}\, \pi_{A'}\, d\pi_{B'} = \pi_{0'}\, d\pi_{1'} - \pi_{1'}\, d\pi_{0'}$$

$$= \Upsilon \lrcorner (d^2Z), \text{ pull-back of projective "volume" 1-form on } \mathbb{P}\tilde{\mathbb{S}}^* ;$$

also $\epsilon^{AB} \dfrac{\partial}{\partial \omega^A} \otimes \dfrac{\partial}{\partial \omega^B}$, which invariantly defines a symplectic 2-form d^4Z/d^2Z on each fibre;

non-holomorphic, conformally invariant

$$\Sigma = Z^\alpha \bar{Z}_\alpha = 2s \quad \text{(twice the helicity)},$$

$$d\Sigma = i\bar{\Phi} - i\Phi \,, \qquad \Theta = d\Phi = d\bar{\Phi} \,;$$

$$\Phi = i\, Z^\alpha d\bar{Z}_\alpha \,;$$

$\Theta = i\, dZ^\alpha \wedge d\bar{Z}_\alpha$, real symplectic structure (Kähler 2-form).

5. Twistor ("first") Quantization

To pass to a quantized theory, we may allow our twistor variables $Z^\alpha_1, \ldots, Z^\alpha_n$, and their complex conjugates $\bar{Z}_{1\alpha}, \ldots, \bar{Z}_{n\alpha}$ to become non-commuting quantum operators, subject to commutation rules:

$$[Z^\alpha_r, Z^\beta_t] = 0 \,, \quad [Z^\alpha_r, \bar{Z}_{t\beta}] = \delta^\alpha_\beta \delta_{rt} \,, \quad [\bar{Z}_{r\alpha}, \bar{Z}_{t\beta}] = 0 \,,$$

taking $\hbar = 1$. This agrees with the standard x^a, p_a quantization when $s = 0$; it is compatible with the twistor symplectic structure; it gives the correct "Poincaré" commutation rules for p^a, M^{ab}; it also gives the SU(2,2) Lie algebra. No factor-ordering problems arise for the Poincaré generators p^a, M^{ab} , but for the helicity operator (for a single twistor system), we now find

$$s = \tfrac{1}{4} (Z^\alpha \bar{Z}_\alpha + \bar{Z}_\alpha Z^\alpha) \,.$$

For a twistor wave-function (for a one-twistor system), in the $\mathcal{T}$-representation we use a function $f(Z^\alpha)$ "independent of $\bar{Z}_\alpha$", i.e. holomorphic in Z^α. Then the "operator Z^α" is represented simply as "multiplication by the twistor variable Z^α" , while the "operator $\bar{Z}_\alpha$" is represented by "$-\partial/\partial Z^\alpha$". Thus, the helicity operator becomes:

$$s = -\tfrac{1}{2} Z^\alpha \partial/\partial Z^\alpha - 1 = -\tfrac{1}{2} \mathcal{T} - 1 \,.$$

In the $\mathcal{T}^*$-representation (i.e. in terms of dual twistor variables) we can use $f(\bar{Z}_\alpha)$ "independent of Z^α" , i.e. anti-

A brief outline of twistor theory. *Cosmology and gravitation; spin, torsion, rotation and supergravity.* Eds P. G. Bergmann and V. de Sabbata, pp. 287–316. Plenum, *New York,* 1980

holomorphic in Z^α. Then, in this case, the "operator Z^α" is represented by "$\partial/\partial\bar{Z}_\alpha$" and the "operator $\bar{Z}_\alpha$" by "multiplication by the variable $\bar{Z}_\alpha$". The helicity operator becomes:

$$s = \frac{1}{2}\bar{Z}_\alpha \frac{\partial}{\partial \bar{Z}_\alpha} + 1 = \frac{1}{2}\bar{\mathbb{T}} + 1 .$$

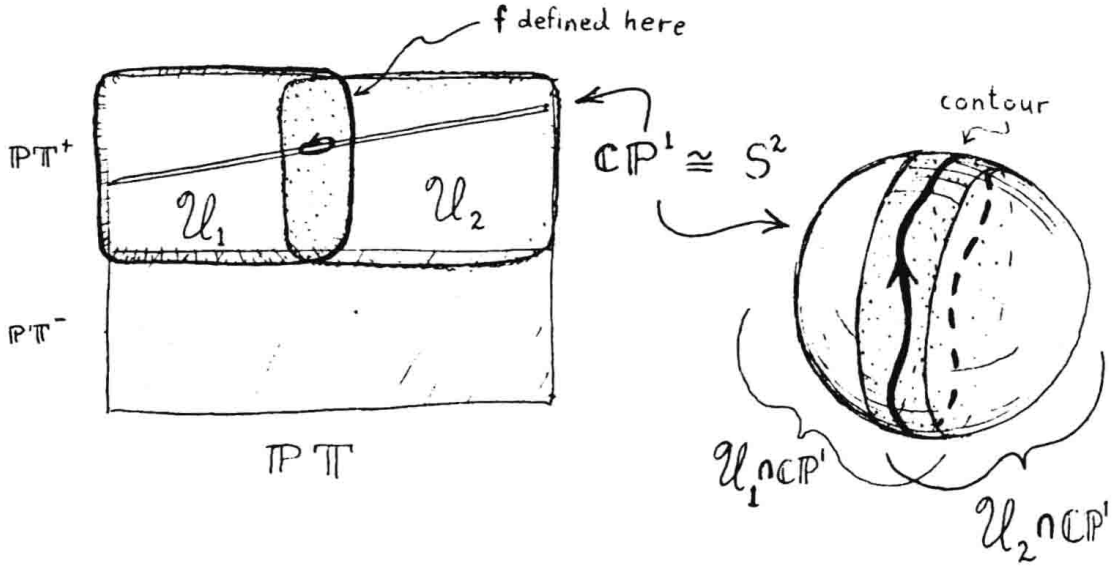
Wave-functions for many-twistor systems are similar. For example, we can take $f(Z^\alpha_r)$ holomorphic. Then we can calculate the operators for m^2, J etc. (In fact $m^2 = 2 \sum_{i<j} I_{\alpha\beta}\, Z^\alpha_i Z^\beta_j\, I^{\rho\sigma}\, \partial^2/\partial Z^\rho_i\, \partial Z^\sigma_j$).

For a 1-twistor system the mass m necessarily vanishes and the helicity is the only scalar quantum number. (See table).

Field	Helicity	$\mathbb{T}$-representation homogeneity	$\mathbb{T}^*$-representation homogeneity
self-dual linear-gravity	+2	-6	2
self-dual Maxwell	+1	-4	0
anti-neutrino	$+\frac{1}{2}$	-3	-1
scalar wave	0	-2	-2
neutrino	$-\frac{1}{2}$	-1	-3
anti-self-dual Maxwell	-1	0	-3
anti-self-dual linear gravity	-2	2	-6

For definiteness, let us stick to the $\mathbb{T}$-representation. For a free wave-function in x-space, we have holomorphicity in the forward tube ($\mathrm{Im}(x^a)$ past-timelike), which corresponds, in $\mathbb{T}$-space, to holomorphicity in $\mathbb{T}^+$ - or in $\mathbb{P}\mathbb{T}^+$. But (apart from polynomials, which are trivial in this context) there are no homogeneous functions that are globally holomorphic

on $\mathbb{T}^+$. What we need is something a little more subtle. We can cover $\mathbb{T}^+$ (or $\mathbb{P}\mathbb{T}^+$) with open sets, say $\mathcal{U}_1$, $\mathcal{U}_2$, so that there exist such functions on the intersection $\mathcal{U}_1 \cap \mathcal{U}_2$. To obtain a space-time wave-function, we perform a contour integral within the complex line ($\mathbb{C}\mathbb{P}^1$) in $\mathbb{P}\mathbb{T}^+$ that represents the point in the forward tube at which the wave-function is to be evaluated.



Then it turns out that the massless free-field equations are automatically satisfied.

Case s = 0 (hom. = −2)

$$\psi(x) = \frac{1}{(2\pi i)^2} \oint f(Z) d^2Z$$

$$= \frac{1}{(2\pi i)^2} \oint f(ix^{AA'}\pi_{A'}, \pi_{A'}) d^2\pi = \frac{1}{2\pi i} \oint f(\ldots) \Delta Z$$

$$\Rightarrow \quad \Box \psi = 0 \ .$$

Case $s = \frac{n}{2} > 0$ (hom. = −n−2)

$$\psi_{A'B'\ldots L'}(x) = \frac{1}{(2\pi i)^2} \oint \pi_{A'}\pi_{B'}\ldots\pi_{L'} f(ix^{AA'}\pi_{A'}, \pi_{A'}) d^2\pi$$

$$\Rightarrow \psi_{A'B'\ldots L'} = \psi_{(A'B'\ldots L')} \ , \quad \nabla^{AA'}\psi_{A'B'\ldots L'} = 0 \ .$$

CHAPTER 120

 A brief outline of twistor theory. *Cosmology and gravitation; spin, torsion, rotation and supergravity.* Eds P. G. Bergmann and V. de Sabbata, pp. 287–316. Plenum, *New York*, 1980

Case $s = -\frac{n}{2} < 0$ (hom. = n−2)

$$\psi_{AB\ldots L}(x) = \frac{1}{(2\pi i)^2}\oint \frac{\partial}{\partial \omega^A}\frac{\partial}{\partial \omega^B}\ldots\frac{\partial}{\partial \omega^L} f(ix^{AA'}\pi_{A'}, \pi_{A'})d^2\pi$$

$$\Rightarrow \quad \psi_{AB\ldots L} = \psi_{(AB..L)} \ , \quad \nabla^{AA'}\psi_{AB\ldots L} = 0$$

The special cases $s = \pm\frac{1}{2}$ give the Dirac-Weyl neutrino equation. When $s = \pm 1$ we get the Maxwell field $F_{ab}(= -F_{ba})$:

$$F_{AA'BB'} = \underbrace{\psi_{AB}\,\epsilon_{A'B'}}_{\text{anti-self-dual (left-handed)}} + \underbrace{\epsilon_{AB}\,\tilde{\psi}_{A'B'}}_{\text{self-dual (right-handed)}}$$

the Maxwell equations $\nabla^a F_{ab} = 0$, $\nabla_{[a}F_{bc]} = 0$ becoming $\nabla^{AA'}\psi_{AB} = 0$, $\nabla^{AA'}\tilde{\psi}_{A'B'}$. When $s = \pm 2$ we get the linearized Einstein curvature field K_{abcd}:

$$K_{AA'BB'CC'DD'} = \underbrace{\Psi_{ABCD}\,\epsilon_{A'B'}\epsilon_{C'D'}}_{\text{anti-self-dual (left-handed)}} + \underbrace{\epsilon_{AB}\epsilon_{CD}\,\tilde{\Psi}_{A'B'C'D'}}_{\text{self-dual (right-handed)}} .$$

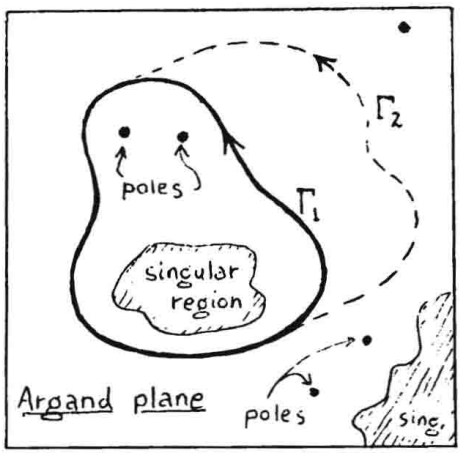
We have $K_{abcd} = K_{[cd][ab]}$, $K_{[abc]d} = 0$, $K^a{}_{bad} = 0$. The "Bianchi identity" $\nabla^a K_{abcd} = 0$, or equivalently $\nabla_{[a}K_{bc]de} = 0$, becomes $\nabla^{AA'}\Psi_{ABCD} = 0$ and $\nabla^{AA'}\Psi_{A'B'C'D'} = 0$.

All these fields enjoy a certain conformal invariance. For rescaling the metric according to $g_{ab} = \Omega^2 g_{ab}$, with $\hat{\epsilon}_{AB} = \Omega\epsilon_{AB}$, $\hat{\epsilon}_{A'B'} = \Omega\epsilon_{A'B'}$, $\hat{\epsilon}^{AB} = \Omega^{-1}\epsilon^{AB}$, $\hat{\epsilon}^{A'B'} = \Omega^{-1}\epsilon^{A'B'}$, we get preservation of the field equations if $\hat{\psi}_{A\ldots D} = \Omega^{-1}\psi_{A\ldots D}$, $\hat{\tilde{\psi}}_{A'\ldots D'} = \Omega^{-1}\tilde{\psi}_{A'\ldots D'}$ i.e. if $\hat{F}_{ab} = F_{ab}$, $\hat{K}_{abcd} = \Omega K_{abcd}$, etc .

6. Contour integrals and sheaf cohomology

To understand the invariance properties of these contour integrals we need some basic sheaf cohomology theory. Consider

first a closed contour integral of a holomorphic function of one variable z. We have:



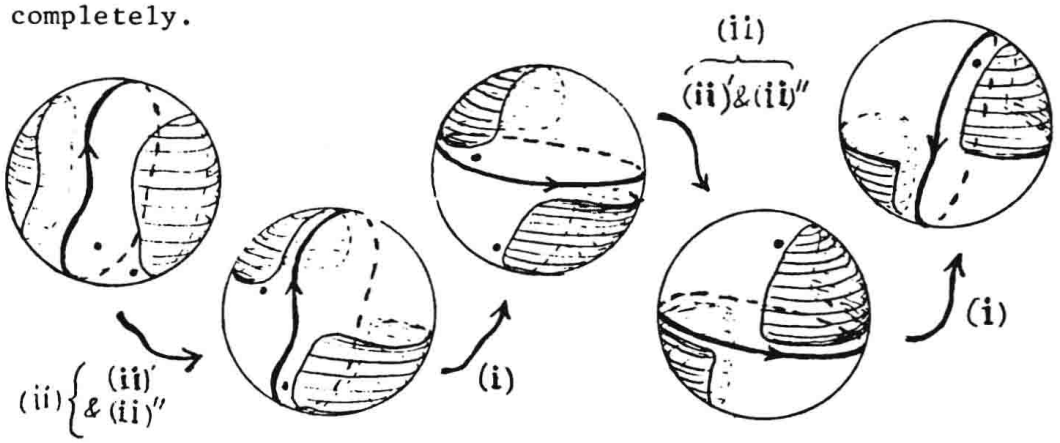
(i) $\oint_{\Gamma_1} f(z)dz = \oint_{\Gamma_2} f(z)dz$

where f is non-singular between Γ_1 and Γ_2 ,

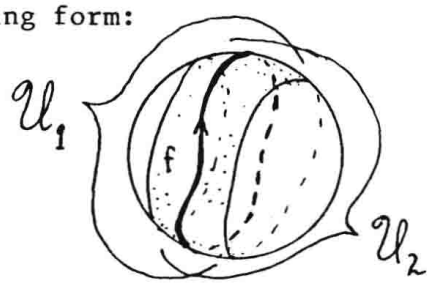
(ii) $\oint_{\Gamma} f(z)dz = \oint_{\Gamma} g(z)dz$

where f - g is non-singular inside Γ .

In the case of the twistor integrals, the Argand plane is replace by the Riemann sphere $\mathbb{C}\mathbb{P}^1$. By repeated applications of (i) and (ii) above, we can change the contour and function completely.



Transformation (ii), on the Rieman sphere S^2, takes the following form:



$f \equiv g$ if $f - g = \overset{\text{(ii)}'}{h_2} - \overset{\text{(ii)}''}{h_1}$

with h_1 regular in $\mathcal{U}_1$, h_2 regular in $\mathcal{U}_2$ where

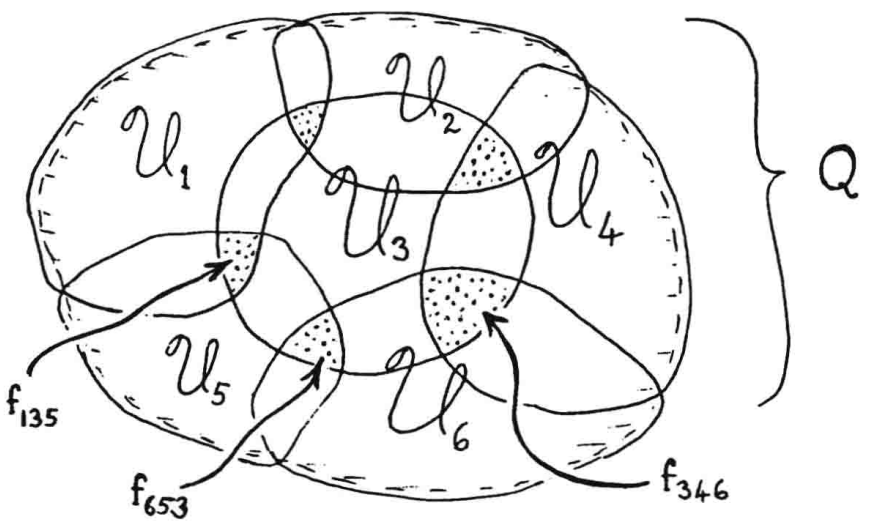
$\mathcal{U}_1$ and $\mathcal{U}_2$ are open sets which together cover S^2 .

But how do we add a function f to a function g, when f and g are defined on different regions? This leads us into (Čech) sheaf cohomology.

Suppose, more generally, that we have some complex n-space Q that is Hausdorff and paracompact. Let $\{\mathcal{U}_i\}$ be an open, locally finite covering of Q. We define a <u>(p-1)-cochain</u> f to be a collection of functions

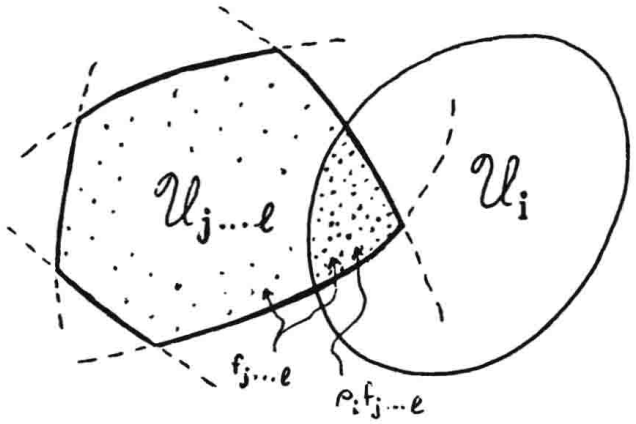
$$\underbrace{f_{ij\ldots\ell}}_{p} = f_{[ij\ldots\ell]}$$

defined on the various p-fold intersections $\mathcal{U}_i \cap \mathcal{U}_j \cap \ldots \cap \mathcal{U}_\ell = \mathcal{U}_{ij\ldots\ell}$



The functions belong to some suitable class (called "sections of a sheaf") defined on open sets of Q. For example, the functions could be variously specified as continuous, or C^k, or holomorphic or they could be local cross-sections of some bundle over Q (e.g. homogeneous functions, regarded as "functions" on $\mathbb{CP}^n$ - "twisted functions"). We define the <u>restriction</u> map ρ_i as follows: for $f_{j\ldots\ell}$ defined on $\mathcal{U}_{j\ldots\ell}$, we write

$\rho_i f_{j\ldots\ell}$ for $f_{j\ldots\ell}$ restricted to $\mathcal{U}_i \cap \mathcal{U}_{j\ldots\ell}$. This satisfies $\rho_i\rho_j = \rho_j\rho_i =: \rho_{ij}$.



The sheaf properties that our funtions must satisfy are:

(i) if $\rho_i f = \rho_i g$ for all i, then $f = g$

(ii) if $\rho_{[i}f_{j]} = 0$ (i.e. $\rho_i f_j = \rho_j f_i$) for all i,j, then $f_i = \rho_i g$ for some g.

Provided that these hold, our "functions" can be very general and, indeed, need not be functions in the ordinary sense - just such that an additive group is defined for each open set.

For cohomology, we need the concept of coboundary. Let $f_{j\ldots\ell} = f_{[j\ldots\ell]}$ be a (p-1)-cochain $f = \{f_{j\ldots\ell}\}$ then the coboundary operator δ maps f to a p-cochain

$$\delta f = \{\rho_{[i}f_{j\ldots\ell]}\}$$

We say that f is a cocycle if $\delta f = 0$ and that f is a coboundary if there is a cochain g such that $f = \delta g$. All coboundaries are cocycles since $\delta^2 = 0$ ($\rho_{[i}\rho_{[j}f_{k\ldots m]]} = 0$ since $\rho_{[i}\rho_{j]} = 0$). Then we define the p^{th} cohomology group, with coefficients in the sheaf $\mathcal{S}$ (class of "functions") to be:

$$H^p_{\{\mathcal{U}_i\}}(Q, \mathcal{S}) = \{\text{p-cocycles}\} / \{\text{p-coboundaries}\} .$$

This is the cohomology group with respect to the particular covering $\{\mathcal{U}_i\}$. To remove the dependence on the covering take finer and finer coverings and form the direct limit.

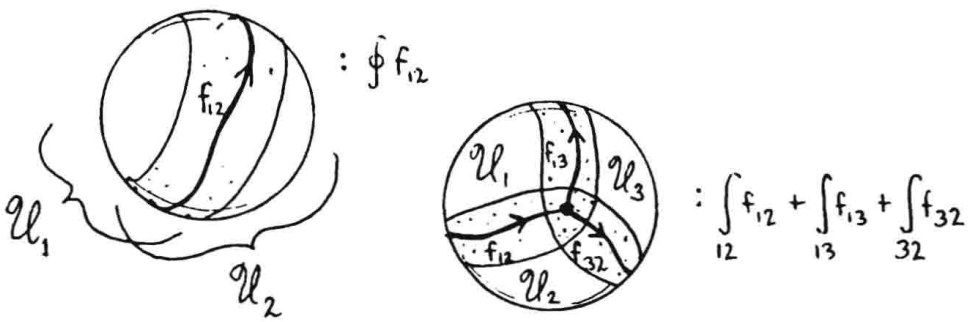
 A brief outline of twistor theory. *Cosmology and gravitation; spin, torsion, rotation and supergravity.* Eds P. G. Bergmann and V. de Sabbata, pp. 287–316. Plenum, *New York,* 1980

This gives the required $H^p(Q,\mathcal{S})$. In practice, however, this limit usually need not be taken, provided that the covering is fine enough. In our case, $\mathcal{S}$ will be a sheaf of holomorphic functions. Then it is sufficient that each $\mathcal{U}_i$ be a Stein manifold (or domain of holomorphy).

In the special case, p = 1, where Q = $\mathbb{CP}^1$ and, possibly, twisted $\mathcal{S}$ = holomorphic functions, two sets $\mathcal{U}_1, \mathcal{U}_2$ covering Q are sufficient. Any holomorphic function f_{12} defined on $\mathcal{U}_1 \cap \mathcal{U}_2$ defines a cochain (with $f_{12} = -f_{12}$, $f_{11} = f_{22} = 0$). A similar function g_{12} defines the same cohomology group element if, for some h_1 holomorphic on $\mathcal{U}_1$ and h_2 holomorphic on $\mathcal{U}_2$, we have

$$f_{12} - g_{12} = \rho_{[1}h_{2]} = \tfrac{1}{2}(\rho_1 h_2 - \rho_2 h_1) \quad .$$

This is just the kind of equivalence that we encountered before, in connection with closed contour integrals on the Riemann sphere. More complicated coverings lead to



branched contour integrals.

Now, it turns out that a twistor wave-function for a massless free field of helicity s is an element of

$$H^1(\mathbb{PT}^+, \mathcal{O}(-2s-2)) \text{ or } H^1(\mathbb{T}^+, \mathcal{O}\{-2s-2\}) \quad .$$

means freely holomorphic but "twisted by -2s-2"	means holomorphic & homogeneous deg. -2s-2 ($\Upsilon f = (-2s-2)f$)

A twistor function $f_{ij}(Z^\alpha)$ is a "representative cocycle" with respect to a covering $\{\mathcal{U}_i\}$ of $\mathbb{PT}^+$ or $\mathbb{T}^+$; $f_{ij} = -f_{ji}$,

$\rho_{[i}f_{jk]} = 0$, and f_{ij} is taken modulo expressions of the form $\rho_{[i}h_{j]}$. The functions f_{ij} are defined on $\mathcal{U}_i \cap \mathcal{U}_j$. The condition $\rho_{[i}f_{jk]} = 0$ is a compatibility requirement on $\mathcal{U}_i \cap \mathcal{U}_j \cap \mathcal{U}_k$ and f_{ij} is taken modulo differences between functions defined on entire $\mathcal{U}_i$'s.

To add two f's defined for different coverings, we need to take a common "refinement". Thus, to form the sum of $\{f_{ij}\}$, for the covering $\{\mathcal{U}_i\}$, and $\{g_{IJ}\}$, for the covering $\{\mathcal{U}'_I\}$, take

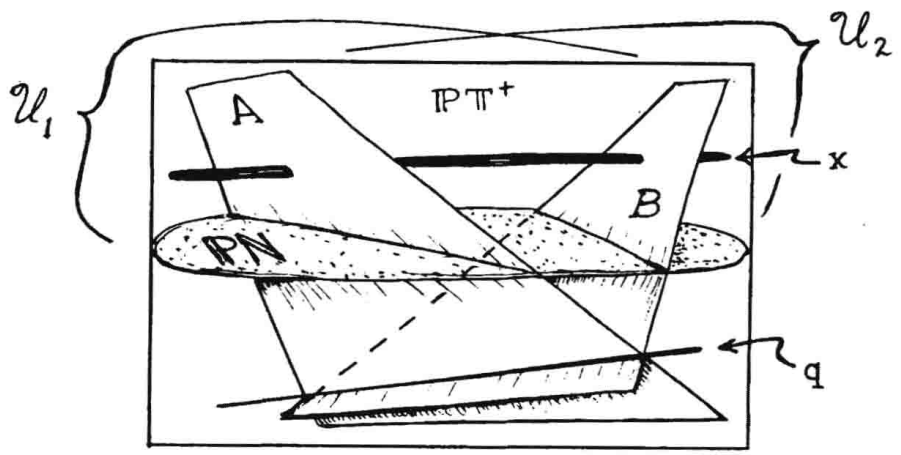
$$\rho_{IJ}\, f_{ij} + \rho_{ij}\, g_{IJ} = F_{iJ,jJ}$$

for the covering $\{\mathcal{U}''_{iI}\}$, with $\mathcal{U}''_{iI} = \mathcal{U}_i \cap \mathcal{U}'_I$ ($\rho_{IJ} = \rho_I\rho_J = \rho_{JI}$, etc.). For a simple example of a twistor function, we can take a scalar wave "elementary state" for which

$$f_{12}(Z) = \frac{1}{A_\alpha Z^\alpha B_\beta Z^\beta}$$

where

$$\mathcal{U}_1 = \mathbb{PT}^+ - B \ , \quad \mathcal{U}_2 = \mathbb{PT}^+ - A \ ,$$



the set B having equation $B_\alpha Z^\alpha = 0$ and A having equation $A_\alpha Z^\alpha = 0$. Writing $A_\alpha = (\alpha_A, a^{A'})$, $B_\alpha = (\beta_A, b^{A'})$ we have, for the space-time field,

$$\psi(x) = \frac{1}{(2\pi i)^2}\oint f_{12}(ix^{AA'}\pi_{A'}, \pi_{A'})d^2\pi$$

$$= \frac{1}{(2\pi i)^2}\oint \frac{d^2\pi}{(i\,\alpha_A x^{AA'}\pi_{A'} + a^{A'}\pi_{A'})(i\,\beta_B x^{BB'}\pi_{B'} + b^{B'}\pi_{B'})}$$

$$= \frac{1}{(i\,\alpha_A x^A_{C'} + a_{C'})(i\,\beta_B x^{BC'} + b^{C'})} = \frac{2}{\alpha^A\beta_A(x^d - q^d)^2}$$

where

$$q^{DD'} = \frac{i\,\alpha^D b^{D'} - i\,\beta^D a^{D'}}{\alpha^A\,\beta_A}$$

This uses

$$\frac{1}{(2\pi i)^2}\oint \frac{d^2\pi}{(\lambda^{A'}\pi_{A'})(\mu^{B'}\pi_{B'})} = \frac{1}{\lambda_{A'}\mu^{A'}}$$

which follows at once from a double application of Cauchy's theorem applied to be variables $\pi_{0'}$ and $\pi_{1'}$. Note that the solution depends only on $B \cap A = q$ (and scaling), not on A and B individually. We can, for example, replace B_α by $\hat{B}_\alpha = B_\alpha + kA_\alpha$ to give

$$g_{12}(Z) = \frac{1}{A_\alpha Z^\alpha\,\hat{B}_\beta Z^\beta},$$

where the covering is now $\{\hat{\mathcal{U}}_1, \mathcal{U}_2\}$ with $\hat{\mathcal{U}}_1 = \mathbb{P}\mathbb{T}^+ - B - \hat{B}$. Then $f_{12} \equiv g_{12}$ since

$$f_{12} - g_{12} = \frac{\hat{B}\cdot Z - B\cdot Z}{A\cdot Z\ B\cdot Z\ \hat{B}\cdot Z} = \frac{k}{B\cdot Z\ \hat{B}\cdot Z}$$

is non-singular on $\hat{\mathcal{U}}_1$.

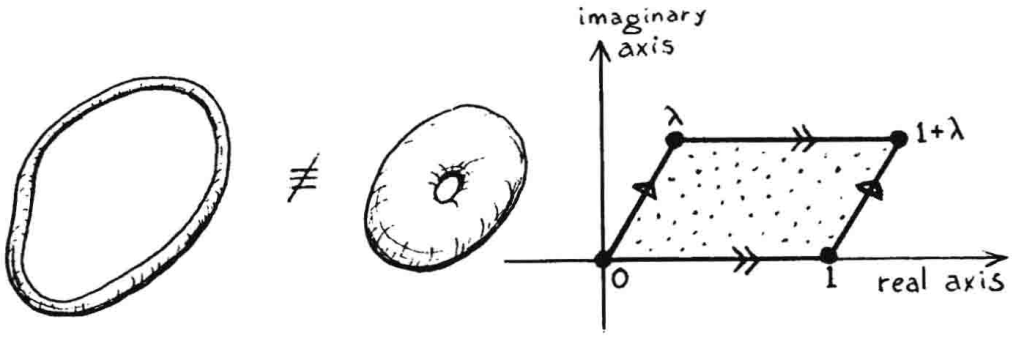
7. Deformations of complex manifolds

As an example, first consider some simple compact complex 1-manifolds.

Case (i) genus 0 : the sphere S^2. All deformations are trivial. There is only one complex sturcture on S^2.

Case (ii) genus 1 : the torus T^2. Now there are non-trivial

deformations, parameterized by 1 complex parameter.

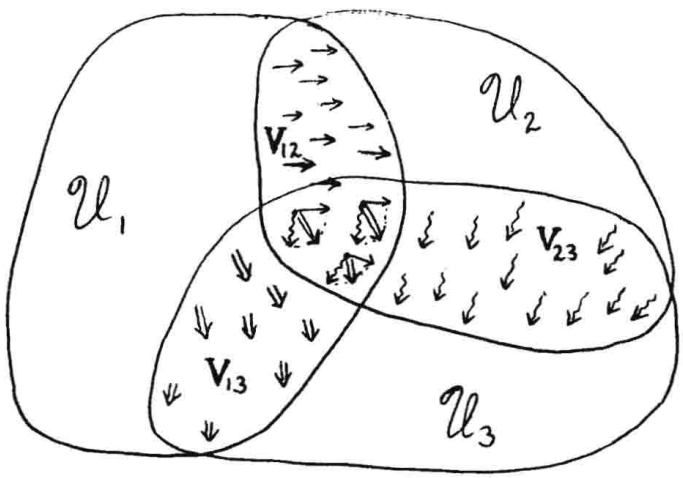


This paramter measures the "shape" of the relevant period parallelogram in the complex plane.

Now consider general complex manifolds. The connection with sheaf cohomology is through infinitesimal deformations. These are described by

$$H^1(Q, \Theta),$$

provided that $H^2(Q, \Theta) = 0$ (the "normal" situation) where Θ denotes the sheaf of holomorphic tangent vector fields on Q. To see this (roughly), let V_{ij} be a tangent vector field on $\mathcal{U}_i \cap \mathcal{U}_j$ and "slide" the patches along the vector fields



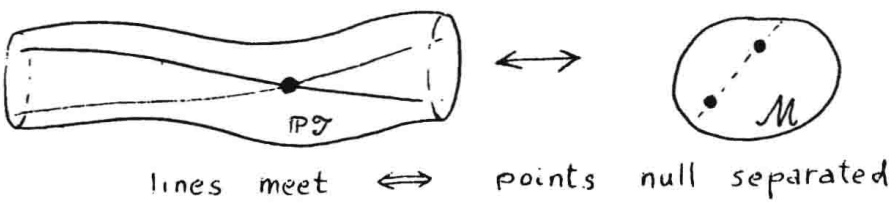
to give new patches. This is "trivial" if $V_{ij} = \rho_{[i}V_{j]}$, since this corresponds simply to sliding the patches in their

 A brief outline of twistor theory. *Cosmology and gravitation; spin, torsion, rotation and supergravity.* Eds P. G. Bergmann and V. de Sabbata, pp. 287–316. Plenum, *New York,* 1980

entirity over Q.

8. Curved Twistor Spaces and Einstein's Equations

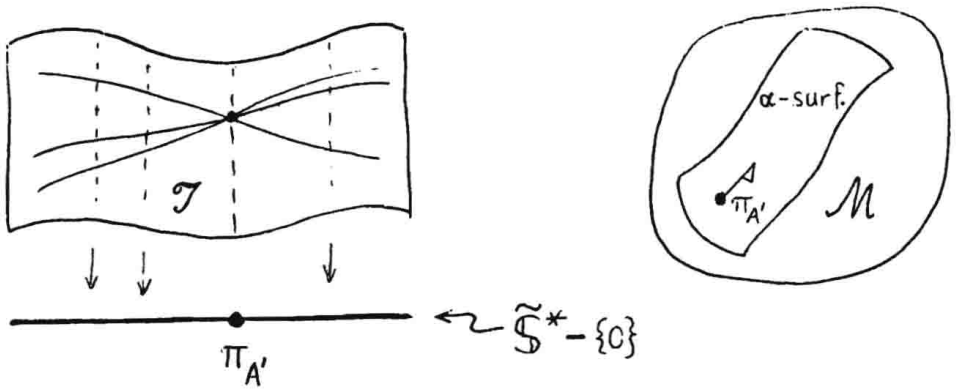
There are no nonltrivial deformations of the whole of $\mathbb{C}\mathbb{P}^3$, but we can deform $\mathbb{P}\mathbb{T}^+$ (or a neighbourhood of a line in $\mathbb{C}\mathbb{P}^3$). Does the corresponding "space-time" still exist? Is there still a suitable 4(-complex)-parameter family of "lines?" These "lines" are to be completely characterized by the fact that they are (i) holomorphic, (ii) compact, and (iii) belong to the correct homology class. Fortunately, there are theorems (due to Kodaira) that say YES (so long as the deformation is not too large). Also, the null-cone structure is necessarily quadratic.



If we fix a point in the (deformed) projective twistor space, we get a 2-parameter family of lines. This is an α-surface. The existence of such a 3-parameter family of α-surfaces implies that the "space-time" is conformally (anti-) self-dual. Thus, in the spinor decomposition of the Weyl curvature tensor C_{abcd}:

$$C_{AA'BB'CC'DD'} = \Psi_{ABCD}\,\varepsilon_{A'B'}\,\varepsilon_{C'D'} + \varepsilon_{AB}\,\varepsilon_{CD}\,\tilde{\Psi}_{A'B'C'D'}$$

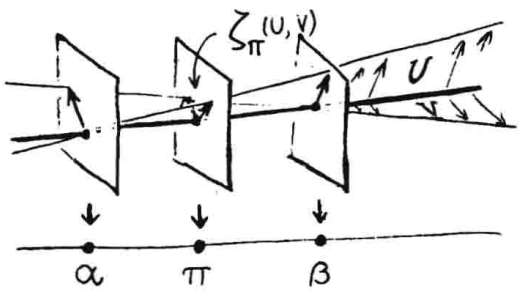
we have $\tilde{\Psi}_{A'B'C'D'} = 0$. The Weyl curvature is anti-self-dual in this sense. For the Ricci tensor (which needs a connection to be defined) also to vanish, we have to have a global parallelism for "primed" spinors. This is achieved by having



a projection from the deformed twistor space $\mathcal{T}$ to $\tilde{\mathbb{S}}^* - \{0\}$ ($\pi_{A'}$-space). Then $\mathcal{T}$ has local holomorphic sturcture identical with that for $\mathbb{T}$, namely:

$$\underbrace{\Upsilon,\ d^4Z,}_{\text{together with projection to } \mathbb{S}^* \text{ suffices to define all remaining structures}}\quad d^2Z,\quad \mathfrak{D}Z,\quad \Delta Z,\quad \zeta = \varepsilon^{AB}\frac{\partial}{\partial\omega^A}\otimes\frac{\partial}{\partial\omega^B}\ .$$

$$\underbrace{\varepsilon^{A'B'}\ \varepsilon^{AB}}_{\text{together give } g^{ab}}$$



$$g(U,V) = \frac{\alpha\cdot\beta}{\alpha\cdot\pi\ \ \beta\cdot\pi}\ \ \zeta_\pi(U,V)$$

for U,V null.

In fact, this construction gives the general anti-self-dual "vacuum space-time" $\mathcal{M}$. The metric for $\mathcal{M}$, and hence also the connection, can be calculated from the structure of $\mathcal{T}$, the quantity $g(U,V) = g_{ab}U^aV^b$ being most easily obtainable when U and V are null vectors in $\mathcal{M}$ (see fig.).

In the case of an infinitesimal deformation of $\mathbb{T}$, we can piece together the space $\mathcal{T}$ out of a number of patches with twistor coordinates Z_i^α, using a twistor function f_{ij}

(representative cocycle for an element of $H^1(\ldots, \mathcal{O}(2))$), so that f_{ij} is homogeneous of degree 2, with $\rho_{[i}f_{jk]} = 0$. Taking ε as infinitesimal ($\varepsilon^2 = 0$) we can express the "patching" as:

$$\rho_j Z^\alpha_i = \rho_i Z^\alpha_j + \varepsilon I^{\alpha\beta} \frac{\partial f_{ij}}{\partial Z^\beta}(Z^\alpha)$$

We can then "exponentiate" such a deformation and pass from linear gravity to non-linear Einstein theory, in the anti-self-dual case.

How about the f_{ij}'s of homogeity degree -6? Can they be correspondingly "exponentiated" to give a <u>self-dual</u> "vacuum"? The problem is that the definition of a point of $\mathcal{M}$ and of its conformal structures, in terms of $\mathcal{T}$, <u>imply</u> that $\mathcal{M}$ is anti-self-dual (because they imply that a 3-parameter family of α-planes exist in $\mathcal{M}$). We need a <u>new</u> definition of space-time point in terms of $\mathcal{T}$ in order to obtain a self-dual "vacuum space-time" $\tilde{\mathcal{M}}$. Now recall our exact sequence expressing Poincaré invariance in the case of flat twistor space $\mathbb{T}$:

$$0 \longrightarrow \mathbb{S} \longrightarrow \mathbb{T} \underset{x}{\longleftarrow\!\!\!\!\!\!\longrightarrow} \tilde{\mathbb{S}}^* \longrightarrow 0 .$$

The original definition for a point of $\mathcal{M}$ may be thought of as a <u>cross-section</u> of the fibration $\mathbb{T} \to \tilde{\mathbb{S}}^*$, i.e. as a map $x : \tilde{\mathbb{S}}^* \to \mathbb{T}$ whose composition with the original projection $\mathbb{T} \to \tilde{\mathbb{S}}^*$ gives the identity on $\tilde{\mathbb{S}}^*$.

$$(\omega^A \pi_{A'}) \longmapsto \pi_{A'}$$

$$(ix^{AB'} \pi_{B'}, \pi_{A'}) \underset{x}{\longmapsto} \pi_{A'}$$

This holds also when $\mathbb{T}$ is deformed to $\mathcal{T}$. In order to construct $\tilde{\mathcal{M}}$, we must try the <u>dual</u> thing (which, equivalently, is the "complex conjugate" thing).

$$0 \longrightarrow \mathbb{S} \underset{\tilde{x}}{\longleftarrow\!\!\!\!\!\!\longrightarrow} \mathbb{T} \longrightarrow \tilde{\mathbb{S}}^* \longrightarrow 0$$

We now interpret a point $\tilde{x}$ (of $\tilde{\mathcal{M}}$) as a map $\tilde{x}: \mathbb{T} \to \mathbb{S}$ which,

when applied to the image of the injection $\mathcal{S} \to \mathbb{T}$ yields the identity on $\mathcal{S}$:

$$\omega^A \longmapsto (\omega^A, 0)$$

$$\omega^A - i\tilde{x}^{AB'}\pi_{B'} \xleftarrow{\tilde{x}} (\omega^A, \pi_{A'}) .$$

For a "deformed" $\mathcal{T}$ and curved $\tilde{\mathcal{M}}$, this requires a suitable regularity condition at I so that the appropriate set of $\tilde{x}$ - maps can be singled out. But I can't be just "regular" in the ordinary sense or the $\tilde{x}$-maps would simply provide a flat $\tilde{\mathcal{M}}$. The most hopeful suggestion, so far, for coding the information of the curvature of $\tilde{\mathcal{M}}$ into the structure of $\mathcal{T}$ appears to be to allow $\mathcal{T}$ to have certain non-Hausdorff properties at I, so as to impose appropriate regularity conditions on the $\tilde{x}$-maps. But the matter remains unresolved, so far.

Now suppose we are given an f_{ij} cocycle on $\mathcal{T}$ (with homogeneity degree ≤ -2), then we can use the x-maps to define a massless field on $\mathcal{M}$. This may be done by just using the old contour integral definition on $x[\tilde{\mathcal{S}}^*]$. The resulting field correctly repsonds to $\mathcal{M}$'s curvature. Can we use $\tilde{x}$-maps to define massless fields on $\tilde{\mathcal{M}}$ (f_{ij} having homogeneity degree ≥ -2)? Apparently we can (assuming the $\tilde{x}$-maps are given). We need a canonical H^1 2-form Ω_{ij} of homogeneity 0 on I ($\mathcal{U}_1$ being $\omega^1 \neq 0$ and $\mathcal{U}_2$ being $\omega^0 \neq 0$):

$$\Omega_{12} = \frac{d\omega^0 \wedge d\omega^1}{\omega^0\,\omega^1} .$$

This pulls back, by $\tilde{x}^{-1}$, to an H^1 2-form, denoted $\Omega_{ij}(\tilde{x})$, on $\mathcal{T}$. Also $\partial/\partial\omega^A$ is well-defined since a flat structure is induced on the constant $\pi_{A'}$ fibres by the $\tilde{x}$-map. We then evaluate

$$\frac{1}{(2\pi i)^4} \oint \frac{\partial}{\partial\omega^A} \cdots \frac{\partial}{\partial\omega^C} f_{[ij} d^2Z \wedge \Omega_{k\ell]}(\tilde{x}) .$$

Assuming that the construction of $\tilde{\mathcal{M}}$ can be carried out

satisfactorily, a natural next step would be to construct vacuum space-times that are neither anti-self-dual nor self-dual (and preferably real). One may anticipate some "average" or "combined" x-map and $\tilde{\mathrm{x}}$-map to yield a general solution of Einstein's vacuum equations. The hope would be that starting (for example) from an asymptotically simple, singularity-free real analytic vacuum space-time $\mathcal{V}$, we could construct its future null infinity $\mathbb{CP}^+$, and then its asymptotic twistor space $\mathcal{T}$. Newman's $\tilde{\mathcal{H}}$-space would then be $\mathcal{M}$ and $\mathcal{H}$-space should be $\tilde{\mathcal{M}}$. The combined construction would be supposed to yield $\mathcal{V}$ back again. Some further ideas would also be needed in order to incorporate sources for the gravitational field.

References and Bibliography

Curtis, G.E., Proc. Roy. Soc. A359 (1978) 133.

Curtis, G.E., Gen. Rel. Grav. 9 (1978) 987.

Curtis, W.D., Lerner, D.E. & Miller, F.R., Gen. Rel. Grav. 10 (1979) 557.

Hansen, R.O. & Newman, E.T., Gen. Rel. Grav. 6 (1975) 361.

Hansen, R.O., Newman, E.T., Penrose, R. & Tod, K.P., Proc. Roy. Soc. A363 (1978) 445.

Hitchin, N.J., Math. Proc. Camb. Phil. Soc. 85 (1979) 465.

Hughston, L.P., Twistors and Particles (Springer Lecture Notes in Physics 97, Springer-Verlag, Berlin 1979).

Hughston, L.P. & Ward, R.S. (Eds.), Advances in Twistor Theory (Pitman Research Notes in Mathwmatics 37, Pitman Press, San Francisco 1979).

Isham, C.J., Penrose, R. & Sciama, D.W., (Eds.), Quantum Gravity: An Oxford Symposium (Clarendon Press, Oxford 1975).

Lerner, D.E. & Sommers, P.D. (Eds.), Complex Manifold Techniques in Theoretical Physics (Pitman Research Notes in Mathematics 32, Pitman Press, San Francisco 1979).

Penrose, R., J. Math. Phys. 8 (1967) 345.

Penrose, R., Int. J. Theor. Phys. 1 (1968) 61.

Penrose, R., J. Math. Phys. 10 (1969) 38.

Penrose, R., Gen. Rel. Grav. 7 (1976) 31.

Penrose, R., Repts. Math. Phys. 12 (1977) 65.

Penrose, R. & MacCallum, M.A.H., Phys. Repts. 6C (1973) 241.

Perjés, Z., Phys. Rev. D11 (1975) 2031.

Perjés, Z., Repts. Math. Phys. 12 (1977) 193.

Tod, K.P., Repts. Math. Phys. 11 (1977) 339.

Tod, K.P. & Perjés, Z., Gen. Rel. Grav. 7 (1976) 903.

Tod, K.P. & Ward, R.S., Proc. Roy. Soc. A368 (1979) 411

Ward, R.S., Phys. Lett. A61 (1977) 81.

Ward, R.S., Proc. Roy. Soc. A363 (1978) 289.

Wells, R.O. Jr., Bull. Amer. Math. Soc. (New Ser.) 1 (1979) 296

Woodhouse, N.M.J., in Group Theoretical Methods in Physics (Eds. Janner, A., Jansen, T. & Boon, M., Springer Lecture Notes in Physics 50, Springer-Verlag, Berlin 1976).

Document generated by Anna’s Archive around 2023-2024 as part of the DuXiu collection (https://annas-blog.org/duxiu-exclusive.html).

Images have been losslessly embedded. Information about the original file can be found in PDF attachments. Some stats (more in the PDF attachments):

{
  "filename": "NDA3ODg1Mzkuemlw",
  "filename_decoded": "40788539.zip",
  "filesize": 120169352,
  "md5": "07d5e56708c9349b0d14a0ed2784bc95",
  "header_md5": "ba8a92de84f02f1b3a36242527b30e05",
  "sha1": "fc5d1b50ab3ac1a109003a56d8a3940ae714c836",
  "sha256": "1b1cb58688e366e75f2dfd9e5bbd5d388579d069fc801e0aa4e8fe00bcccf8c5",
  "crc32": 1184462795,
  "zip_password": "",
  "uncompressed_size": 150995293,
  "pdg_dir_name": "ROGER PENROSE COLLECTED WORKS VOLUME 3 1976-1980_40788539",
  "pdg_main_pages_found": 693,
  "pdg_main_pages_max": 693,
  "total_pages": 718,
  "total_pixels": 4413022536,
  "pdf_generation_missing_pages": false
}